Reconstitution par arbres de régression du rayonnement visible descendant horaire sur la France continentale, à partir de données in situ et de simulations : Spatialisation et vérification sur des données indépendantes

D. Brion, Jean-Christophe Calvet, P. Le Moigne, B. Ghattas, Florence Habets

METEO-FRANCE

CENTRE NATIONAL
DE RECHERCHES METEOROLOGIQUES

NOTE DE TRAVAIL
DU GROUPE DE METEOROLOGIE A MOYENNE ECHELLE

N°82

RECONSTITUTION PAR ARBRES DE REGRESSION DU RAYONNEMENT VISIBLE DESCENDANT HORAIRE SUR LA FRANCE CONTINENTALE, A PARTIR DE DONNEES IN SITU ET DE SIMULATIONS : SPATIALISATION ET VERIFICATION SUR DES DONNEES INDEPENDANTES

par

D. BRION, J.-C. CALVET, P. LE MOIGNE, B. GHATTAS, F. HABETS

OCTOBRE 2005

Table des matières

Résumé

La nécessité de disposer sur toute la France de séries horaires de rayonnement visible descendant a mené à une reconstitution en deux étapes, à partir de paramètres facilement disponibles :

- Une reconstitution en des points où sont disponibles la durée d'insolation, éventuellement des résultats de simulations Safran et des observations de nature et d'étendue des couches nuageuses. La méthode de reconstitution employée est basée sur des moyennes d'ensembles d'arbres de régression. Cette méthode permet de prendre en compte les non linéarités entre les divers intrants météorologiques. Elle est comparée avec d'autres ajustements non-linéaires. Les critères de choix entre les divers algorithmes statistiques sont : la qualité des résultats, la rapidité des apprentissages et la facilité des maintenances logicielles. On n'a pas rencontré de contradiction entre ces 3 critères. Des liaisons sont d'abord établies entre le rayonnement horaire et les autres paramètres météorologiques sur toutes les stations disposant de l'intégralité des données nécessaires. Elles sont d'abord testées par des méthodes de cross-validation sur ces mêmes stations, puis appliquées en tous les points disposant de mesure simultanées de durée d'insolation et de nébulosité (environ un par département). Ces pseudo-observations supplémentaires viennent compléter le réseau d'observations de rayonnement horaire, dont la répartition spatiale est très irrégulière.
- La spatialisation à l'échelle de Safran (maille de 8×8km) est alors effectuée par krigeage ordinaire.

La validation de ces traitements a porté sur 11 stations indépendantes, non gérées par Météo-France, et sur une période (2004) n'ayant servi à aucun réglage.

Chapitre 1

Introduction

1.1 Généralités

Le rayonnement visible est un des termes du bilan d'énergie, et, à ce titre, il est important de disposer de valeurs réalistes de ce paramètre météorologique en toute station. Cependant, le nombre de capteurs de rayonnement est relativement faible, du moins en ce qui concerne les stations météorologiques à observation humaine, et leur qualité inégale. Ceci est d'autant plus gênant que tous les autres termes du bilan d'énergie et d'eau y sont mesurés dans de bonnes conditions. Ces termes peuvent, aussi, être reconstitués par des formules étalonnées sur une grande variété de situations (cas de l'infrarouge descendant). Ce n'est pas tout à fait le cas pour le rayonnement visible, qui est reconstitué en routine à l'aide de formules d'Angstroem (1924), valides pour une durée mensuelle ou, à la rigueur, décadaire. Canellas, Merlier et Pérarnaud (1994) ont identifié et vérifié[1] une relation linéaire valable pour toute la France métropolitaine, saison par saison, entre les durées d'insolation et les rayonnements décadaires. La période d'identification était comprise entre 1971 et 1992, pour 10 stations mesurant simultanément ces deux paramètres. Ces formules ont une application opérationnelle. Soler (1990) a comparé, sur des critères d'écart quadratique et de biais, 7 formulations du type Angstroem en 77 stations de mesure européennes, sur un an de données mensuelles. La seule formule tenant compte de l'altitude donnait les résultats les plus mauvais, et la meilleure formule était une formule relativement simple, dont les coefficients dépendaient du mois. Ces formules de type Angstroem sont basées sur la connaissance de la durée d'insolation cumulée. Cependant, des équations de

[1] C'est à dire : tabulé les erreurs de reconstitution (biais, écart absolu moyen), station par station, la vérification portant , vu le grand nombre de données et le très faible nombre de paramètres, sur le fichier d'apprentissage.

régression différentes peuvent être utilisées suivant la saison (cas de DP/SERV/Agro), permettant ainsi de distinguer les situations de ciel clair de printemps et d'automne grâce à l'intensité du rayonnement direct. Si on veut travailler à des échelles temporelles plus fines que la décade, ce qui peut se produire pour des applications en temps réel (par exemple, un modèle de bilan hydrique et d'énergie comme ISBA est alimenté de préférence avec des séries horaires), il faut trouver d'autres formules empiriques ou d'autres moyens de reconstitution.

Parmi les autres moyens de reconstitution, on a employé :

- une formule disponible dans [OMM, 1987] liant le rayonnement solaire horaire reçu par ciel non nuageux à la tension de vapeur au sol, à l'altitude et à la hauteur du Soleil; cette formule est recommandée pour la vérification des étalonnages et aurait une précision de 20%; elle pourrait être améliorée en tenant compte du trouble atmosphérique, déduit de la couleur du ciel, mais ceci n'a pas été mis en oeuvre faute d'archives; un substitut basé sur une fonction linéaire de la visibilité horizontale n'a pas introduit d'amélioration. Il faut rappeler que cette formule n'est utilisable que par ciel clair, ce qui n'est pas très fréquent sous nos climats.

- le rayonnement visible reconstitué par Safran, décrit plus loin.

Disposant de 2 façons de reconstituer des valeurs de rayonnement solaire, de sophistications très différentes, il semble prioritaire de se concentrer sur des reconstitutions basées sur des statistiques. En effet, on dispose de quelques séries d'observations simultanées de rayonnement et d'autres grandeurs météorologiques plus usitées telles que la durée d'insolation et la couverture nuageuse utiles pour une reconstitution, et, une fois la liaison entre ces paramètres établie lors d'une phase d'identification, pratiquement toutes les stations à observation humaine peuvent disposer de rayonnement reconstitué.

On verra successivement :

- Les méthodes statistiques employées pour établir une liaison empirique entre le rayonnement et des paramètres météorologiques plus facilement disponibles en une station à observation humaine.
- Les performances des reconstitutions ponctuelles sur des fichiers les plus indépendants possibles, d'abord sur un très petit nombre de stations où on peut interpréter les résultats, puis sur toutes les stations à observation humaine mesurant aussi le rayonnement visible, en s'intéressant surtout aux résultats

globaux.

- Sachant qu'il y a au moins une station à observation humaine par département, les techniques de reconstitution précédemment décrites permettent de compléter un réseau irrégulier par des observations synthétiques en une cinquantaine de stations. Ainsi, les techniques de krigeage permettent de remplir complètement les zones où il n'y a pas d'observations. Le principe et un exemple d'application en seront présentés sommairement,
- La validation finale est effectuée sur des postes gérés par un autre organisme que Météo-France, pour l'année 2004 qui n'a servi, ni à identifier des liaisons entre des observations humaines et le rayonnement, ni à caler les paramètres de la fonction de structure utilisée en spatialisation.

On va tout d'abord présenter les méthodes d'identification que l'on a pu comparer et mettre en oeuvre pour relier le rayonnement en un point à diverses variables explicatives (arbres de régression et polspline), le nombre forcément réduit de logiciels d'identification utilisés découle de contraintes détaillées en Annexe 1.

Remarque: la source de données principale étant la BDCLIM, on a choisi d'exprimer les résultats dans les unités de cette dernière : cumul de rayonnement horaire en J cm^{-2} h^{-1}, position des points en hectomètres (hm).

1.2 Arbres[2]

1.2.1 Principe informatique et éléments de chronométrage

La liaison entre la grandeur à reconstituer et les p variables explicatives continues est représentée par une combinaison linéaire d'indicatrices de polytopes de R^p dont les cotés sont parallèles aux axes de coordonnées et constituant une partition de R^p construits séquentiellement par segmentation : pour chaque variable, on définit un seuil optimal qui permet d'affecter 2 valeurs à la variable à expliquer; on choisit la meilleure variable (au sens de la variance intraclasse de la grandeur à expliquer) et on recommence dans chaque classe. La mise à jour des variances intraclasse et des moyennes est incrémentale, l'exploration exhaustive (dans les logiciels classiques) des coupures implique de réordonner chaque variable explicative. Cette découpe n'est pas affectée par une

[2] la mise en oeuvre de la méthode des analogues avec 2 variables explicatives donne presque (remplacer médiane par moyenne dans le traitement des doublons) le même résultat que les arbres, qui, de par leur mécanisme de sélection, s'accommodent de davantage de variables explicatives.

transformation monotone sur une variable.

Dans le cas de variables qualitatives comme par exemple la nature des nuages bas ou le temps présent, codées sans que leur ordre n'ait de sens, la notion d' intervalle disparaît et tous les sous-ensembles possibles sont essayés dans les logiciels anciens; cette complexité exponentielle avec le nombre de modalités est actuellement souvent limitée par recodage par la moyenne, conditionnelle à la modalité, de la grandeur à prévoir (ex : pour des nuages bas stratiformes, la valeur à prévoir est V1, pour des stratocumuli, la valeur à prévoir est V2, et V3 pour des cumuli; la nature des nuages est remplacée par les Vi), ceci à chaque étape de la construction de l'arbre.

Les temps d'exécution lors de l'apprentissage sont convexes en la taille N de l'échantillon d'apprentissage (dans le cas très favorable d'un arbre de profondeur 1, le temps est celui d'un tri, proportionnel à p*N*log(N)) dans le cas de variables explicatives quantitatives, et linéaires en N (toujours pour un arbre de profondeur 1 (du type si X1< seuil, alors affecter V1 à y, sinon, affecter V2)), mais peuvent croître exponentiellement avec le nombre de modalités possibles pour les variables qualitatives ; cet argument permet de ne pas envisager d'utiliser sans précautions comme variable explicative le temps présent, codé sur 100 valeurs, si on dispose d'un logiciel ancien; certains logiciels plus modernes sont, quant à eux, limités par la représentation interne (codage disjonctif sur 32 bits, typiquement, dans randomForests ou tree) des variables qualitatives ...
Si la croissance d'un arbre et la fabrication de sous arbres emboîtés n'est limitée que par l'effectif minimal pour calculer une moyenne, et si cette croissance se fait de façon très déséquilibrée, de sorte que, à chaque fois, on ait une branche à effectif minimal, la profondeur de l'arbre sera proportionnelle à l'effectif initial N, ce qui induit un temps d'exécution croissant comme N*N*ln(N) dans un cas très défavorable. En pratique, les essais effectués indiquent une croissance quadratique avec la taille du problème[3].

1.2.2 Disponibilité et quelques applications.

Ces algorithmes ont été mis au point par Breiman, Friedman, Olshen et Stone en 1984, d'autres techniques de construction d'arbres existent mais cette technique est souvent implémentée à des fins de comparaison. Loh (2002) critique la méthode de sélection des variables explicatives exposée ci dessus, et le logiciel 'guide' met en oeuvre des algorithmes de sélection de variables moins gloutons (binaires gratuits

[3] les logiciels commerciaux revendiquent parfois une croissance linéaire du temps d'identification avec la taille; ceci est obtenu en répartissant entre le minimum et le maximum de chaque variable explicative un nombre fixé de coupures, ce qui dispense de l'étape de tri au prix éventuel d'une dégradation

disponibles pour Linux et Windows, et intégré à SPSS).
Les résultats que nous exposerons ci-après font appel à ***rpart*** de Therneau et Atkinson, greffé à ***R*** (un cousin gratuit de ***Splus***, conçu par Ihaka &Gentleman, 1997, installable sur tout calculateur moderne prévu pour des traitements interactifs). Il s'agit d'une extension de ***R*** qui a été jugée, en 2002, suffisamment fiable informatiquement (par 2 experts indépendants) et suffisamment intéressante statistiquement pour être incluse d'office dans toute installation de ***R*** standard.

Les applications météorologiques sont rares, ce qui est dû au fait que les méthodes linéaires sont souvent suffisantes en prévision et peuvent, de par le nombre de stations concernées, nécessiter des temps d'identification déjà très conséquents; ces applications sont essentiellement liées, pour ce qui est de la prévision, à des phénomènes assez peu linéaires tels que la neige de lac (Burrows 1991) et la détection de plafonds bas en Californie à partir d'observations satellitaires et de sorties de modèle (Blankert et al., 2004). En matière de lien entre la pollution et la météorologie, la littérature est beaucoup plus abondante et ne sera pas évoquée exhaustivement : Ryan (1995) a employé des arbres classiques et des ajustements linéaires pour la prévision à courte et moyenne échéance de l'ozone à Baltimore ; Ghattas (1999) a comparé des arbres comme ceux que nous pouvons employer et des arbres stabilisés par moyennes d'ensemble pour la prévision à 12 heures d'échéance de pointes d'ozone dans les Bouches du Rhône. Ces algorithmes ont fait l'objet d'applications opérationnelles par Airparif pendant 4 ans (Bel & al., 1999) pour la prévision d'ozone et de dioxyde d'azote à Paris, et continuent d'être utilisés en routine par Airmaraix.

A titre d'exemple, la Figure 1 montre un petit arbre de régression pour reconstituer, pour une gamme de hauteurs de soleil fixée, le rayonnement visible : les principaux paramètres sont, dans ce cas, la nébulosité totale et le fait que, par forte nébulosité totale, les nuages hauts soient absents ou invisibles (ceci correspond à une prédominance des nuages moyens ou bas).

1.2.3 Quelques défauts et leurs remèdes.

Le temps d'identification est long, mais, pour des problèmes portant sur peu de stations et un seul paramètre à prévoir, ceci est d'autant moins gênant que l'on peut l'estimer et s'organiser en conséquence.
Prévoir une grandeur continue sous forme de somme finie d'indicatrices de partitions de R^p peut théoriquement introduire des cassures intolérables; dans le cas du rayonnement visible horaire, déjà discrétisé et dont les cumuls servent à faire évoluer les bilans d'énergie, ceci a été jugé a priori négligeable sauf peut-être pour

représenter les valeurs les plus extrêmes (cette référence à l'utilisation finale impliquerait, en toute logique, que, non seulement la reconstitution, mais aussi l'ensemble de la chaîne reconstitution + utilisation soit validé).
Les variables cycliques (direction du vent, mois) ne sont pas gérées en tant que telles; peu de logiciels statistiques gratuits le font, ***guide*** est une exception.
Le point le plus gênant réside dans l'instabilité des arbres isolés, qui peut se manifester, lors d'une petite perturbation des fichiers d'apprentissage, par de profondes modifications dans la structure des arbres voire des performances sur fichier test. Breiman 1994 et Ghattas 1999, en montrent quelques exemples et y remédient par la moyenne de plusieurs arbres obtenus par ré-échantillonnage (par tirage avec remise à probabilités constantes) du fichier d'apprentissage. Le nombre d'arbres intervenant dans cette procédure est de l'ordre de 10 à 25. Le défaut de cette procédure de bootstrap est éventuellement son appel à des tirages aléatoires, on verra plus loin comment y pallier.

1.3 Polspline

Il existe beaucoup d'autres algorithmes statistiques (Hastie, Tibshirani & Friedman, 2001). Stone et al. 1997, modélisent une grandeur continue comme combinaison linéaire de fonctions de base 'splines' linéaires par morceaux en une seule variable (ou produits de 2 telles fonctions), avec les contraintes suivantes :

- Un produit de 2 fonctions ne peut exister que simultanément à ces 2 fonctions.
- Une cassure ne peut être appliquée à une variable que si cette variable intervient déjà sous forme linéaire classique.

Ces contraintes facilitent l'interprétation des modèles obtenus et accélèrent leur identification. Le temps d'apprentissage est difficile à évaluer, du fait de butées (sur le nombre de fonctions de base avant sélection et le nombre de cassures possibles) tout à fait raisonnables pour les grands effectifs (il n'en reste pas moins long et surtout imprévisible, car dépendant de sélections), et cette procédure est assez flexible pour se ramener si besoin à une régression linéaire classique (on peut en effet interdire toute cassure et tout produit) munie d'une étape de sélection progressive ascendante puis de sélection descendante. Il en découlerait donc une instabilité, comme pour la suite (sélection de sous ensembles de variables explicatives + régression) évoquée par Breiman 1994, et par Buhlmann et Yu 2002 ; par contre, ces derniers auteurs démontrent aussi la stabilité des régressions par splines, si bien qu'on ne peut pas avoir d'idée a priori sur la stabilité dans le cas général.
L'interprétation découle de l'analogie avec la régression linéaire. La recherche

d'éventuelles non linéarités est simplifiée par la possibilité de tracer les isolignes de la grandeur à prévoir dans un système d'axes définis par 2 variables explicatives à la discrétion de l'utilisateur, les autres variables étant constantes (c'est une application de formule sans chercher à savoir si le point de fonctionnement est réaliste et cohérent avec les variations que l'on contrôle; la nouvelle version ***polspline*** ne tracerait des isolignes que dans la portion du plan où il y a effectivement des valeurs).
polspline peut prendre en compte des valeurs continues ou qualitatives. Cependant, il est sensible à des variables explicatives très corrélées linéairement, au même titre que la régression linéaire.

1.4 Safran.

Ce modèle d'analyse météorologique au voisinage du sol a été à l'origine conçu par Durand et al. (1993) pour fournir des intrants spatialisés au modèle d'évolution du manteau neigeux Crocus. Il fournit des séries horaires de température, humidité et vent au voisinage du sol, par interpolation optimale, ainsi que les rayonnements solaire et de grande longueur d'onde par fabrication, dans chaque zone Symposium éventuellement scindée en fonction du relief, d'un profil atmosphérique type, déduit des analyses Arpege ou CEPMMT par interpolation spatio-temporelle (entre 0000, 0600, 1200 et 1800 UTC) et des nébulosités alimentant les calculs radiatifs de Ritter & Geleyn (1992).

D'abord limité à la prévision des avalanches en zone de montagne, il a été couplé au modèle ISBA (Noilhan & Planton, 1989) en mode forcé afin d'établir les bilans d'énergie et d'eau à l'interface entre le sol, l'air et la végétation. Cette interface entre le sol, la biosphère et l'atmosphère est à son tour couplée au modèle hydrologique MODCOU (Ledoux & al., 1989) qui reconstitue le débit des rivières avec beaucoup de réalisme dans le cas du bassin du Rhône (Golaz-Cavazzi, 1999; Etchevers, 2000; Boone & al., 2002) et du bassin Adour-Garonne (Morel, 2002). Les résultats satisfaisants et d'autres applications potentielles ont amené l'extension du domaine géographique de SAFRAN à toute la France métropolitaine (Le Moigne 2002), sur des carrés de 8 km de côté. Le domaine temporel s'étend de août 1995 à nos jours, du fait de son passage en opérationnel fin juillet 2002, ce qui en fait une très longue série homogène quant à la méthode de calcul. A noter qu'il est très rare que des séries de données analysées aient simultanément une telle finesse géographique et une telle profondeur temporelle tout en gardant les mêmes calculs, ce qui peut être nécessaire pour des besoins d'homogénéité.
Le Moigne, 2002 a décrit les qualités de la reconstitution des paramètres interpolés par Safran; pour la température, le vent, l'humidité, voire la pluie, où

plus de 200 points de vérification étaient disponibles, l'accord dépasse largement ce qui est habituel avec des données prévues. Le rayonnement infrarouge descendant a été vérifié à l'aide des stations de Carpentras et du Col de Porte, les deux points de mesure du rayonnement atmosphérique disponibles en France et des corrections tenant compte de la nébulosité et de l'altitude ont été proposées.

Le rayonnement visible a été vérifié à l'aide de 6 stations, dont cinq d'excellente qualité, et situées à des altitudes très différentes dans le bassin versant du Rhône ou en son voisinage. Là aussi, des corrections en fonction de l'altitude et de la nébulosité ont été identifiées. Ces corrections étaient plus faibles en valeur relative que pour le rayonnement de grande longueur d'onde, qui intervient 24h sur 24 dans les bilans d'énergie. Cette réserve faite, il n'en reste pas moins une disproportion entre le nombre de stations employées pour vérifier les paramètres météorologiques courants et celles ayant servi à vérifier le rayonnement visible. La capacité de Safran à reconstituer le rayonnement visible sera donc vérifiée dans le Chapitre 3. Pour donner une idée de ce que l'on peut faire de mieux sans Safran, le Chapitre 2 présentera des résultats obtenus avec des mesures météorologiques de routine pour des stations à observation humaine : il y en a plus de 100 en France, dont moins de 30% disposent de mesures de rayonnement horaire.

Chapitre 2

Reconstitution dans le Nord de la France à l'aide de la durée d'insolation horaire

On commencera par rappeler une méthode des analogues, qui a servi à faire les premières reconstitutions. Les difficultés inhérentes à cette méthode ont incité à se tourner vers des logiciels à diffusion plus large susceptibles de traiter simplement davantage de variables explicatives. Le choix de ***rpart*** plutôt que ***polspline*** découlait alors ... de la supériorité, légère mais constante, du premier logiciel sur des données de validation croisée; ceci, ajouté au fait que les temps d'apprentissage de ***polspline*** étaient longs dans notre cas, a forcé la décision.

2.1 Rappel et critique d'une méthode des analogues.

Elle a été mise au point pour permettre, à partir de la durée d'insolation horaire, de la nébulosité totale et de la pluie, de reconstituer le rayonnement dans les stations synoptiques de la DIRIC pour lesquelles on souhaite effectuer, grâce à ISBA, un bilan d'énergie et d'eau. Deux techniques de reconstruction extrêmement simples ont été écartées d'emblée :

- calculer une valeur moyenne régionale au même moment, solution susceptible de mener à trois types de défauts :
  - une heure où il pleut localement peut disposer d'un rayonnement reconstitué suffisamment fort pour perturber les calculs d'ETR,
  - les mesures de rayonnement dans les stations voisines, si elles sont entachées d'erreurs, peuvent corrompre une moyenne régionale,
  - l'évolution prévisible de réseaux soumis à rationalisation (aboutissant à

terme à un réseau moins dense mais de qualité homogène) rend absurde toute tentative de validation.

- prendre la valeur en la station la plus proche géographiquement, au même moment, offre exactement les mêmes défauts, qui peuvent être aggravés par le fait que les réseaux actuels de l'Ile de France sont plus denses que ceux du Centre, où les distances peuvent dépasser 80 km.

On a donc préféré se concentrer sur les paramètres météorologiques mesurés sur place, et on pouvait, soit faire appel à des formules déjà disponibles, soit identifier une relation statistique sur des stations ayant le même climat et mesurant de façon fiable le rayonnement, et reporter informatiquement cette relation en la station où on souhaite disposer de valeurs de rayonnement. On a éliminé les formules toutes faites, souvent valides à l'échelle du mois (Black & al., 1954) ou de la journée (Hunt & al., 1998) du fait des besoins spécifiques de données horaires et aussi parce que transposer des formules est une source de fautes de transcription, d'unités ou de domaines de validité. Par exemple, Black et al. ont identifié 6 formules d'Angstroem, dépendant de la climatologie des épaisseurs optiques des nuages et signalent qu'elles dépendent aussi de la technologie employée pour mesurer la durée d'insolation. Enfin, le grand nombre de ces formules semblerait de nature à accroître encore la confusion.

On a été amené à choisir une méthode des analogues pour deux raisons :

- la nature de la liaison éventuelle entre le rayonnement et la durée d'insolation horaire est a priori mal connue, elle est très simple à coder informatiquement.
- il s'agit d'une méthode non paramétrique inspirée de celle des plus proches voisins décrite, pour la discrimination, dans der Megreditchian 1993 et dont le principe est le suivant :
    - l'historique des observations du paramètre à reconstituer et des variables explicatives correspondantes est conservé dans une table,
    - on définit une distance entre 2 vecteurs de variables explicatives, classiquement une somme pondérée à coefficients positifs des carrés des différences des valeurs de chaque variable,
    - on choisit comme valeur à reconstituer celle qui correspond au minimum des distances et, en cas d'ambiguïté (fréquent), on affecte la moyenne ou la médiane de toutes les valeurs possibles.

Il n'y a pas besoin d'identification si on connaît la forme et les paramètres de la distance. Cette méthode est assez lente lors d'une mise en oeuvre opérationnelle (obligation de balayer l'intégralité d'une table) mais elle est stable. A la différence

de la régression, une erreur, même arbitrairement grande, sur une valeur de la grandeur à prévoir dans les fichiers historiques ne se répercute pas sur la forme d'une équation, ce qui est une forme de résistance. Zorita & von Storch (1999) signalent aussi que l'optimisation des coefficients de la distance est très difficile, du fait de nombreux minima mais que c'est la méthode qui respecte le mieux les valeurs extrêmes (ce qui était fort intéressant dans leur problème de recherche des effets d'une dérive du climat sur la fréquence des pluies fortes). Par contre, l'interprétation en est jugée très pauvre.

En pratique, on fabrique des tables heure par heure et pour chaque mois ou décade en 5 stations (Bourges, Reims, Rennes, Tours et Trappes) mesurant simultanément le rayonnement visible et la durée d'insolation et situées, soit dans la DIRIC, soit suffisamment près pour que les variations de l'heure solaire ne dépassent pas 10 minutes, et que celles de latitude soient petites (la hauteur du soleil varie vite d'un jour à l'autre au voisinage des équinoxes, ceci incite à découper en décades les mois de mars, avril, septembre et octobre).
Les observations utilisées s'étalent entre 1994 et 2001, période où les technologies d'observation de la durée d'insolation et du rayonnement n'ont pas varié. Dans le cas général, ces tables mensuelles ont une longueur de 5 (stations)× 8 (années) × 30 (jours) et présentent 61 valeurs possibles d'insolation et 10 valeurs possibles de nébulosité totale, pour 90 % d'heures sans pluie. Il y aura donc des valeurs équivoisines, et la médiane des reconstitutions possibles est alors choisie, du fait de sa résistance à d'éventuelles valeurs aberrantes dans les tables d'analogues (celles qui ont pu être détectées ont été retirées, induisant une très faible amélioration).

Cette découpe en heure et mois permet de se placer à heure solaire constante et de limiter les temps de validation : en effet, on a procédé comme Hunt et al. en retirant une station des tables d'analogues et en comparant les valeurs de rayonnement observées en cette station aux valeurs de rayonnement reconstituées avec des tables d'analogues réduites. Cette opération de validation étant quadratique avec la profondeur historique, son temps d'exécution est divisé par plus de 12. Par contre, on n'a pas envisagé d'optimiser les paramètres de la distance, parce que l'on ne connaît que des méthodes d'exploration systématique. Si maintenant on envisageait de rajouter un paramètre, une telle optimisation deviendrait vraisemblablement nécessaire mais les temps de paramétrage ont été jugés prohibitifs sans changement de logiciel ni de calculateur. Dans notre cas de deux variables explicatives (la pluie est rare à l'échelle horaire) discrétisées et de longues séries d'analogues, rechercher l'optimum de la forme des distances n'a pas de solution unique dès que les tables sont complètement remplies.

Les résultats mettaient en évidence un accroissement généralisé de la qualité des reconstitutions à partir de 1996, attribué à la mise en place centralisée d'une politique de maintenance préventive (renvoi régulier du capteur et de son électronique en atelier pour re-conditionnement éventuel). Les biais, écarts absolus moyens, coefficients de corrélation et de Spearman et coefficients de corrélation de tendance étaient cohérents entre eux et indiquaient une légère supériorité de cette reconstitution en été par rapport à l'hiver (d'octobre à mars)[4]. Les coefficients de corrélation pour les heures où le rayonnement visible prédomine sont largement supérieurs à ce que Hunt et al. trouvaient pour des reconstitutions à échelle journalière. Ceci, compte tenu de la lourdeur d'une éventuelle procédure d'optimisation, a servi de critère d'arrêt en faisant implicitement les hypothèses que reconstituer des séries horaires puis les cumuler à l'échelle du jour présente moins de défauts que de reconstituer directement des séries quotidiennes et que les coefficients de corrélation valides sur le Sud du Canada peuvent se comparer à des coefficients de corrélation dans le Nord de la France...

Un autre critère d'arrêt était la supériorité vis à vis d'une moyenne régionale (moyenne des rayonnements horaires mesurés à moins de 80 km de Trappes, au voisinage duquel est disponible un réseau assez dense de points de mesure). La satisfaction de ce critère ne permet cependant pas de comprendre si la supériorité d'une reconstitution avec des paramètres disponibles sur place découle de la grande variabilité géographique du rayonnement horaire ou de la mauvaise qualité des observations voisines.

Outre l'impossibilité d'interpréter finement ces résultats, des défauts liés au faible nombre de variables explicatives utilisées (durée d'insolation, nébulosité totale et logarithme de la pluie) ne pouvaient pas être résolus :

- il existe des heures où le ciel est suffisamment voilé pour qu'il n'y ait pas d'ombres portées mais où le rayonnement est plus fort que sous un cumulonimbus, par exemple. Cette information sur la nature des nuages existe, mais n'est pas prise en compte.
- le rajout de la pluie apportait un gain faible mais constant de performances; on ne sait pas si cela est lié à des problèmes instrumentaux (dépôts d'eau sur les optiques) signalés par Le Moigne ou à une réalité physique.

On peut aussi remarquer que cette façon de procéder par découpe en heure est très difficile à généraliser si les stations utilisées s'étendent davantage géographiquement. Par ailleurs, une structure d'arbre dont la croissance n'est pas limitée par des butées fonctionnerait pratiquement de la même manière (à

[4] l'inverse aurait été gênant

l'exception du remplacement de la médiane par la moyenne) que la méthode du plus proche voisin si les seules variables explicatives étaient la durée d'insolation et la nébulosité totale. S'il y a lieu d'utiliser d'autres variables explicatives, les mécanismes de sélection des logiciels d'arbres les mettront en évidence.

2.2 Comparaison Polspline/Arbres.

On est donc incité à essayer d'autres logiciels,
avec la liste des prédicteurs suivants :

- la durée d'insolation cumulée sur la même heure que le rayonnement visible à reconstituer,
- la visibilité, seuillée à 20 kilomètres pour pouvoir généraliser à des stations mal dégagées comme Trappes, en zone périurbaine (ce seuillage dégrade très légèrement des stations bien dégagées comme Dijon, et améliore légèrement la reconstitution à Trappes),
- les codes définissant la nature des nuages bas, moyens et hauts.

Ces cinq paramètres sont estimés subjectivement à la fin de l'heure sur laquelle portent des paramètres cumulés. On a veillé à ce que ni la visibilité, ni la nébulosité totale ne manquent : dans cette configuration (très fréquente), le codage de la nature des nuages élevés, par exemple, comme observation manquante est alors dépourvu d'ambiguïté et indique que l'observateur était présent mais qu'une couche de nuages moyens ou bas empêchait l'identification des nuages supérieurs.
La quantité de précipitation (via son logarithme, ce codage est sans conséquence pour des logiciels d'arbres invariants par transformations monotones des intrants, et améliore les reconstitutions par splines de ***polspline***), la tension de vapeur d'eau, la couverture des nuages bas, le mois, l'heure et le cosinus de l'angle solaire zénithal (tel qu'il est calculé) complètent cette liste de variables explicatives potentielles.

Pour simplifier la validation et pouvoir étendre le nombre de stations utilisables, on a abandonné une découpe en mois et heure au profit d'une découpe en station et cosinus de l'angle solaire zénithal. Le seul réglage du logiciel d'arbre qui ne soit pas classique consistait à tolérer une croissance des arbres seulement limitée par le nombre minimal d'éléments, et non par un critère statistique obtenu par validation croisée, assez chronophage; ce réglage est courant pour des moyennes d'ensemble d'arbres (Breimann 1994), et on a admis que le fait, en reconstitution, de moyenner les valeurs après application des arbres obtenus pour chaque station avait un effet stabilisateur équivalent au 'bagging'. L'autre inconvénient réside dans le grand nombre d'arbres extrêmement complexes, au point que l'interprétation en est très longue. On n'a pas cherché à vérifier tous les arbres, sur toute leur profondeur à

cause de leur grand nombre : ils sont donc restés en l'état.
Cependant, on peut déjà remarquer que la durée d'insolation est le paramètre qui suscite systématiquement la première coupure; son fort potentiel explicatif est connu depuis plus de 80 ans.

Les identifications par splines ont utilisé la même liste de prédicteurs que pour les arbres[5], et on s'est aperçu que la découpe en cosinus de l'angle solaire zénithal était superflue, et on a donc travaillé seulement station par station. Les critères de comparaison étaient :

- le coefficient de Nash, part de la variance (exprimée en pour 1000) de la grandeur à reconstruire expliquée par la reconstitution. Il est asymétrique : par exemple, si la grandeur à reconstruire est mesurée avec un capteur bloqué (ne serait-ce que par la présence d'ombres très fortes), on ne peut pas le calculer. Il est à la fois sensible au bruit et aux biais,
- le biais (écart entre le cumul de rayonnement recalculé et observé).

On a comparé dans le Tableau 1 les résultats obtenus avec des reconstitutions par arbres, obtenues dans le pire des cas (reconstitution avec des arbres identifiés sur 4 stations différentes de la station considérée, avec une découpe en peu d'angles solaires zénithaux) et les résultats obtenus par ***polspline*** dans le meilleur des cas (reconstitution avec 6 autres stations, avec fixation automatique, lors de l'apprentissage, d'un paramètre de régularisation) pour Bourges, Saint-Quentin et Reims, stations affectées par peu de problèmes de mesure. Les stations supplémentaires dont les performances ne sont pas affichées pour des raisons de simplicité sont Trappes, Tours, Rennes et Dijon. Ce tableau porte sur des cumuls quotidiens, les données brutes horaires sont un peu moins bien reconstituées si on considère les coefficients de corrélation (non disponibles ici). Les scores sont calculés pour des données telles qu'elles ont été extraites de la BDCLIM, sans nettoyage préalable, alors que les identifications ont porté sur des données partiellement débarrassées de valeurs douteuses. La faiblesse de ce filtrage vient du fait qu'il est essentiellement subjectif et qu'il pouvait favoriser des mécanismes déjà connus, tout en inhibant la sélection de nouvelles variables explicatives.

[5] pour des raisons de simplicité, on a écarté d'autres indicateurs de performance, tels que le coefficient de corrélation des rangs, un indicateur robuste mais invariant par toute transformation monotone d'une des deux grandeurs comparées; il serait intéressant si les grandeurs étaient susceptibles d'atteindre des valeurs anormalement grandes, ce qui n'est pas le cas du rayonnement. De même, le coefficient de corrélation de tendance (corrélation entre les variations d'un jour à l'autre des deux séries) offrirait l'avantage de s'affranchir partiellement du caractère saisonnier du rayonnement visible, qui explique plus de 50% de sa variance, et de s'affranchir aussi de son cycle diurne. On a préféré s'affranchir du cycle diurne en ne considérant que des cumuls quotidiens, et ne pas introduire d'interprétations optimistes liées à la forte saisonnalité en calculant ces critères de qualité pour des saisons de 3 mois

STATION	ARBRES	POLSPLINE (meilleur des cas)
Bourges Hiv.	979:-21	976:-9
Bourges Pri.	977: -16	977: 7
Bourges Eté	974: -8	978: 1
Bourges Aut.	968:-22	955:-28
St.Que. Hiv.	959:9	955:13
St.Que. Pri	960:26	952:46
St.Que. Été	964:14	959:17
St.Que. Aut	973:3	966:-6
Reims Hiv.	973:-10	970: -6
Reims Pri.	978:-24	975:-15
Reims Eté	974:-15	971:-24
Reims Aut.	974:-11	966:-18

TABLEAU 1 - Reconstitution des cumuls quotidiens de rayonnement par la méthode des arbres et par Polspline : scores sur 3 stations (Nash : biais). Nash désigne le coefficient de Nash (mutiplié par 1000), biais est la différence (en J cm^{-2} h^{-1}) entre la valeur reconstituée et la valeur observée.

Les biais et indicateurs de qualité affichés ne présentent aucun caractère saisonnier marqué, ceci résulte peut-être de l'effort qui a été fait de désaisonnaliser par la hauteur du soleil. Le fait que les arbres de ***rpart*** soient légèrement, mais quasi systématiquement, supérieurs aux reconstitutions par splines ***polspline*** est difficile à expliquer, et n'est valable que pour le rayonnement (pour un paramètre non borné, tel que le vent, les logiciels d'arbres peuvent mener, sans précaution d'emploi, à des résultats gênants en cas de tempête). On peut éventuellement l'attribuer à une meilleure connaissance des mécanismes de stabilisation pour ***rpart***. Une autre explication résiderait dans le fait que les interactions entre variables météorologiques utiles seraient plus compliquées qu'un produit d'au maximum 2 fonctions linéaires par morceaux d'une variable explicative.

2.3 Importance des variables explicatives.

On a vu ci-dessus que la nouvelle liste des variables explicatives mène, pour deux techniques d'identification, à des résultats ayant un faible potentiel d'amélioration. Par contre, la critique appliquée à la méthode des analogues d'être une boîte noire qui n'élucide pas l'utilité éventuelle d'une variable s'applique alors aussi aux

ensembles d'arbres (alors qu'un des avantages revendiqués des arbres de régression découlerait de leur facilité d'interprétation) par l'accroissement du nombre de variables explicatives et par le nombre d'arbres utilisés. Pour y remédier, il convient de revenir sur le cas où une variable explicative est manquante :
Lors de la phase d'identification, une fois la meilleure variable et la meilleure cassure trouvées, ***rpart*** recherche la séquence des autres variables et des cassures correspondantes qui respectent au mieux la découpe optimale précédente, ceci pour tenter de reproduire la structure initiale d'arbre en aval de cette découpe. Cette opération prend le même temps que de déterminer la meilleure coupure si le nombre de variables explicatives est grand. Il est donc tout à fait possible de s'accommoder de valeurs manquantes, si leur définition est sans ambiguïté, au prix d'un doublement du temps d'identification et de l'usage de la mémoire. On peut alors, pour chaque variable, calculer la variance que l'on expliquerait si seule cette variable était présente et définir ainsi une hiérarchie de variables de moins en moins importantes. Ainsi, B. Ghattas 1999 a-t-il pu sélectionner 3 variables parmi 20 variables explicatives initiales, alors que deux de ces trois variables n'apparaissaient pas dans l'arbre de départ. Cet ordonnancement des variables explicatives est bien plus rapide qu'une sélection progressive.

L'ordonnancement obtenu pour le rayonnement est peu étonnant : la nébulosité totale et l'insolation sont des prédicteurs très informatifs, le détail des couvertures nuageuses venant en complément. Parmi ce détail, la nature des nuages hauts est la plus informative. La visibilité est plus explicative que la quantité de précipitations horaires. Les autres variables explicatives sont de peu d'intérêt, ce qui est conforme à ce que l'on attendait.

2.4 Conclusion.

On peut reconstituer avec beaucoup de réalisme le rayonnement en une station à observation humaine disposant de mesures de durées d'insolation horaire; la meilleure méthode, tant du point de vue des performances qu'en temps d'exécution, consiste à désaisonnaliser les fichiers d'apprentissage (et aussi, à faire les essais dans les mêmes conditions) pour des logiciels d'arbres par découpe en petites classes (effectif de l'ordre de 1000 à 3000) de cosinus de l'angle solaire zénithal, et, pour chaque petite classe, de travailler station d'apprentissage par station d'apprentissage. Ceci permet de :

- ménager l'avenir informatique si la mesure du rayonnement en une station est déclarée fausse ou si on veut ajouter une station,
- garantir des temps d'apprentissage raisonnables (environ 5 minutes par station pour un apprentissage sur la période 1995-2002, le test et divers

diagnostics graphiques sur des valeurs quotidiennes),

- valider ces choix.

Le rayonnement horaire est alors reconstitué par moyenne, dans la tranche de hauteur du soleil adéquate, des résultats de prévision par les arbres de régression identifiés pour chaque station en excluant naturellement la station qui sert à valider cette méthode. Pour que cette reconstitution garde de bonnes performances, il est nécessaire que la durée d'insolation et les observations humaines soient présentes. Une autre limitation découle des conditions d'apprentissage, sur des stations de plaine à basse altitude où les éventuels effets d'atténuation par les aérosols et la vapeur d'eau varient peu, alors qu'il faudrait introduire une correction fonction de l'altitude dans le cas général.

Chapitre 3

Reconstitution à l'aide du rayonnement visible calculé par Safran

3.1 Résultats à 1200 UTC

Il s'agit d'une heure synoptique où sont disponibles à la fois les analyses d'Arpège ou du CEPMMT et les observations, sans que des interpolations temporelles des profils verticaux nécessaires aux calculs de rayonnement pour les autres heures de la journée ne compliquent la description. De plus, d'éventuels masques ou rideaux d'arbres sont peu susceptibles de perturber les mesures servant de validation. On peut trouver un grand nombre de stations mesurant ou ayant mesuré le rayonnement, et donc susceptibles de servir à vérifier le bon comportement de Safran. On s'est donc limité à une quarantaine de stations de plaine, au Nord de 45.5N et offrant les particularités suivantes :

- en dehors de la DIRIC, seules les stations faisant l'objet d'une documentation informatisée RADOME[6] et jugées a priori suffisamment bien implantées pour mesurer le rayonnement sans trop de perturbations liées aux bizarreries du site ont été conservées,
- dans la DIRIC, pour des raisons de contingence, toutes les stations ont été passées en revue si elles étaient en service en 2003. Ce biais défavorable invalide naturellement d'éventuelles comparaisons avec les autres directions interrégionales.

La Figure 2 détaille sous forme de boîtes à moustaches (« boxplots ») le

[6] les documentations Radome, disponibles sous forme informatique et centralisées depuis décembre 2002, ne portent que sur la qualité d'un site de mesure et non sur le respect éventuel des consignes d'étalonnage d'un capteur. Le fait de se limiter à des stations bien situées devrait faciliter, à terme, la gestion de la maintenance et de son historique.

comportement des cumuls sur le mois, entre et 11 h et 12h, observés et reconstitués par Safran, ceci pour les stations et les mois où le coefficient de Nash entre l'observation et la reconstitution était positif ou nul. Ce dernier cas correspond à un mois où la reconstitution triviale par la moyenne mensuelle, si on pouvait l'observer, mènerait à la même erreur quadratique. Les rares mois où le coefficient de Nash était négatif peuvent s'interpréter de 2 façons :

- la mesure est juste, mais Safran est affecté de défauts; c'est le cas classique en vérification de prévision, où l'on peut s'appuyer sur la qualité de paramètres contrôlés rigoureusement en routine, ce contrôle étant effectué par une source indépendante,
- des erreurs de mesure privent de sens la comparaison, ceci pouvant être amplifié par des blocages. L'exemple type est celui d'une station sur un chantier de construction qui présentait des cycles diurnes extrêmement faibles, et peu réalistes (normalement, le contrôle automatique, basé essentiellement sur des seuils horaires, ne peut pas détecter ce genre de défauts).

On retrouve dans la majorité des stations la sous estimation par Safran signalée pour d'autres stations par Le Moigne. Les deux exceptions parmi les stations à observation humaine sont Nancy, et Dijon, où une forte fréquence de nuages bas l'hiver nécessiterait (si cette explication est pertinente) un dépouillement saisonnier ou fonction de la nature de la couverture nuageuse.

La Figure 3 indique, pour chaque année, l'évolution des cumuls mensuels des énergies visibles observées et reconstituées pour 6 stations : quatre présentent des cumuls observés légèrement supérieurs aux cumuls Safran, avec peu de variation d'une année à l'autre; Nancy sous-estimerait le rayonnement, et Trappes souffre d'un défaut d'étalonnage en 2001, qui a été confirmé par deux personnes indépendantes[7]. Naturellement, on a veillé à ce que les périodes incriminées ne servent pas à des identifications statistiques (mais elles restent utilisées pour la vérification).

3.3 Résultats toutes heures confondues

Le fait que Safran sous-estime un peu le rayonnement ne se retrouve pas de façon aussi nette pour des cumuls mensuels, toutes heures confondues; ceci peut être lié à trois causes:

[7] ceci est suffisamment rare pour qu'on conserve précieusement ce genre de défauts, pour pouvoir éventuellement tester la capacité de les détecter

- la non prise en compte de la pluie et des condensations, ces dernières pouvant se produire tôt le matin ou tard le soir,
- l'interpolation temporelle des profils verticaux, sur une base de 6 heures,
- la présence de nuages bas, pouvant être mal diagnostiqués à partir d'un profil vertical.

Par contre, même si on n'arrive pas actuellement à expliquer ces différences de comportement, on obtient de meilleures corrélations sur des cumuls mensuels que si on se limite aux données de 12h.

3.4 Arbres de régression basés sur Safran combiné avec des observations humaines

Cette dernière remarque, jointe au fait que certaines stations aéronautiques telles que Roissy ou Orly ne disposent pas d'un capteur de durée d'insolation et aussi à la constatation que les capteurs de durée d'insolation, quoique bien connus, ne sont pas dispensés de pannes ni de dérives, incite naturellement à s'inspirer de ce qui a été fait plus haut en matière de reconstitution par arbres pour des stations à observation humaine. La façon la plus immédiate consiste à réutiliser la même découpe en fonction du cosinus de l'angle solaire zénithal et la même liste de paramètres, en substituant simplement le rayonnement Safran à la durée d'insolation. Le seul développement supplémentaire a consisté à comparer les déterminations de l'importance des variables par substitution systématique de chaque variable à une autre formulation de l'importance, disponibles dans 'RandomForests, feb. 2003' de Breiman, déterminée par brassage aléatoire de la colonne dont on veut déterminer l'utilité relative, cette opération étant répétée quelques centaines de fois. Comme aucune incohérence flagrante n'a pu être détectée, on a gardé la précédente version déterministe du calcul de l'importance d'une variable.

Deux cas ont été traités, l'un se limitant à 1200 UTC, pour essayer de dégager des variables complémentaires, et le cas général, pour aboutir à une reconstitution du rayonnement dans tous les cas possibles pour les stations à observation humaine.

Le cas de la reconstitution des cumuls de rayonnement à 1200 UTC est présenté dans le Tableau 2, et mis en regard de la reconstitution utilisant seulement le rayonnement Safran à cette heure qui ne fait pas l'objet d'une interpolation.

STATION	ARBRES	SAFRAN seul
Bourges Hiv.	858: -4:25:92	548:-14:45:75
Bourges Pri.	759:-11:43:87	305:-37:73:72
Bourges Eté	755: -2:39:86	377:-23:63:69
Bourges Aut.	805: -3:24:89	479: -9:40:73
St.Que. Hiv.	836: 3:24:89	678: 2:32:79
St.Que. Pri	718: 0:45:84	374:-17:67:67
St.Que. Été	744: -1:38:85	443:-17:56:72
St.Que. Aut	841: 0:20:90	736: -3:26:86
Reims Hiv.	848: 0:23:90	500: 1:43:69
Reims Pri.	693:-11:48:83	212:-22:77:60
Reims Eté	742: -5:40:85	290:-21:66:65
Reims Aut.	832: 0:21:90	635: -4:31:75

TABLEAU 2 - Reconstitution des cumuls de rayonnement à 1200 UTC par la méthode des arbres de régression et par Safran seul : scores sur 3 stations (Nash : biais : EQM : corrélation). Nash désigne le coefficient de Nash (mutiplié par 1000), biais est la différence (en J cm^{-2} h^{-1}) entre la valeur reconstituée et la valeur observée, EQM est l'erreur quadratique moyenne et la corrélation est le coefficient de corrélation des rangs (multiplié par 100).

On voit que la part de variance expliquée est considérablement accrue par l'ajout d'autres variables explicatives, quelle que soit la saison. Le fait que, l'hiver (janvier à mars) et l'automne, Safran soit supérieur aux autres saisons au moins pour des scores insensibles aux changements d'unités doit cependant être signalé.

La Figure 4 permet de visualiser l'importance des variables utilisées. Le poids très fort du rayonnement théorique en ciel clair est lié à l'absence de dé-saisonnalisation pour les calculs d'arbres à 1200 UTC. Le rôle très grand des couches de nuages moyens et élevés a déjà été évoqué, ainsi que l'apport vraisemblablement faible de l'observation des cumuls de pluies horaires.

La Figure 5, décrivant l'importance des variables désaisonnalisées par découpe en fonction de la hauteur du soleil met en évidence le rôle prépondérant des couches nuageuses, les grandeurs présentant une variation saisonnière devenant moins utiles.

Du fait du grand nombre (au moins du point de vue du temps d'apprentissage informatique) de stations employées pour construire des arbres de régression dans le Nord de la France (près de 10, et il faut multiplier par 20 découpes en hauteur du

soleil, et que les temps d'entraînement pour faire un arbre sont de quelques ordres de grandeur plus élevés que pour une régression), il est naturel de s'inquiéter de la lourdeur de cette procédure et de regarder quel peut être l'apport d'une station, soit en la supprimant (sélection progressive descendante), soit en la rajoutant à partir de rien (première étape d'une sélection progressive ascendante). La Figure 6 montre que les résultats obtenus avec des arbres identifiés avec une seule station sont nettement plus mauvais que ceux obtenus avec 9 stations, où on constate une diminution de l'erreur de prévision de 30% environ. Cette diminution de l'erreur est beaucoup plus forte que le fait de préférer une station à une autre, la station la moins préférable étant Nancy, qui, une fois isolée, mène à des arbres dont l'application aux autres stations est affectée du maximum d'erreurs et dont la suppression provoque le moins d'erreurs possible. L'apport de Tours, que ce soit individuellement ou insérée dans un ensemble d'arbres est aussi positif. Par contre Trappes, pris isolément, apporte beaucoup, mais sa suppression d'un ensemble dégrade peu la qualité de la prévision correspondante, ceci pouvant s'expliquer par un contrôle particulièrement sévère du rayonnement à Trappes. Des identifications calées sur ces données peuvent expliquer grossièrement les données des autres stations, mais des particularités censurées à tort lors d'un contrôle subjectif cessent d'être reproductibles.

Une fois cumulés sur un jour et surtout sur un mois, les rayonnements visibles reconstitués à l'aide de Safran combiné avec des d'observations humaines offrent des performances très voisines de celles d'une reconstitution basée sur la durée d'insolation et des observations humaines.

3.4 Conclusion pour Safran dans le Nord de la France.

La sous estimation du rayonnement visible de Safran signalée par Le Moigne a pu être vérifiée sur d'autres stations que celles qui lui ont servi à valider Safran, à 1200 UTC, la description des heures non synoptiques étant plus complexe.
Une fois cumulé à l'échelle du jour et surtout du mois, le rayonnement de Safran se corrèle très bien à l'observation, ceci permet d'envisager son utilisation pour reconstituer une observation manquante, ou comme source indépendante et homogène pour détecter d'éventuelles dérives (les capteurs de durée d'insolation ou de rayonnement ne sont pas toujours présents, ni exempts de défauts).
L'apport de l'observation humaine combinée avec Safran permet d'aboutir à des reconstitutions presque aussi réalistes à l'échelle horaire que ce que l'on obtient par traitement non linéaire de l'insolation et des particularités du couvert nuageux visible. Ce traitement composite a cependant pour inconvénient de faire appel à deux sources de données distinctes, ce qui peut induire une certaine baisse de la

fiabilité si on envisage une application en temps réel, ou, dans le meilleur des cas, une gestion assez compliquée des horaires. Il n'est pas justifié par le nombre de stations supplémentaires qui, du fait de l'absence de mesures de durée d'insolation ne peuvent faire l'objet d'estimations du rayonnement par reconstitution comme expliqué dans le Chapitre 2.

Il reste cependant à signaler que la procédure de validation employée que ce soit dans les Chapitres 2 et 3, consistant à retirer une station d'apprentissage tout en travaillant sur la même période, peut induire un léger biais d'optimisme si on envisage d'utiliser ces méthodes sur de nouvelles périodes, du fait de la corrélation géographique des variables explicatives. Le choix de cette procédure de validation découlait de raisons historiques (comparaison avec les résultats de Hunt) et de sa simplicité.

En toute rigueur, si on envisage une utilisation routinière, il faudrait prévoir une mise à jour tout aussi routinière des identifications, ce qui est recommandé pour une adaptation statistique, du moins dans un contexte de modifications technologiques. Cette éventuelle mise à jour a été simplifiée en prenant en compte la croissance quadratique du temps d'identification avec la longueur des fichiers d'apprentissage. On a aussi constaté, ce qui est rare, que cette optimisation du temps informatique se traduisait par un gain de performances.

Naturellement, ces reconstitutions ne peuvent être envisagées que pour des stations de plaine (en France, elles conduiraient à rajouter plus de 50 points où une valeur de rayonnement horaire serait disponible); pour des stations de montagne, des identifications paramétriques sont établies (Le Moigne, 2002) et pleinement justifiées par leur simplicité et la prise en compte physique des phénomènes.

Par ailleurs, des listes de variables explicatives utiles ont pu être établies, mettant en évidence le rôle connu de la nature de la couverture nuageuse.

Chapitre 4

Généralisation sur toute la France continentale

Les méthodes consistant à combiner Safran et des observations humaines développées ci-dessus pour le Nord de la France ont été généralisées à l'ensemble de la France (triplement du nombre de points d'apprentissage).
Cette généralisation a consisté essentiellement à veiller à maîtriser les listes de variables explicatives, de façon à pouvoir éclaircir le rôle de chacune de ces variables. Dans le même temps, une méthode de validation alternative au retrait d'une station de la liste des stations sur lesquelles on cale un modèle statistique (la station retirée pouvant servir de test) a été utilisée. Accessoirement, les logiciels ont été simplifiés et accélérés.

4.1 Validation tenant compte des liaisons spatio-temporelles.

Dans les essais portant sur le Nord de la France, la technique de validation consistait à retirer une station de la liste des stations servant à l'identification, et à établir des statistiques d'erreurs sur cette station pour les modèles statistiques calés sur le reste des stations; en réitérant cette opération, il était possible de disposer de statistiques d'erreurs pour toutes les stations. La grosse faiblesse de cette façon de faire résidait dans le fait que les données météorologiques sont corrélées géographiquement, si bien qu'on ne peut pas savoir si des performances affichées sont réalistes ou si elles sont systématiquement optimistes.

On a choisi de garder cette découpe en stations, aidant l'interprétation si besoin,

mais de retirer 2 années glissantes lors de l'identification à des fins de calculs des performances (c'est-à-dire empêcher qu'elles servent simultanément à l'identification d'arbres et au test, même si ces 2 opérations portent sur des stations disjointes). Le choix de cette période de 2 années glissantes sur laquelle portent les tests, indépendante de la période d'identification, résulte d'un compromis entre la profondeur de l'archivage jugée utile pour pouvoir disposer de tous les codes de nuages et le temps de cette validation croisée.
Les statistiques d'erreurs sont alors établies sur les 2 années retirées pour la station retirée (méthode employée par Blankert et al. 2005).

Un dernier point reste à noter : on a choisi de ne pas faire intervenir lors des apprentissages les stations ayant une implantation non recommandée et les stations ayant eu des problèmes gênants dont on a pu prendre connaissance. Ce filtrage, forcément non exhaustif, permet aussi de diminuer les temps d'apprentissage et de validation croisée. Naturellement, ces stations servent quand même à l'établissement de statistiques d'erreurs.

4.2 Résultats avec toutes les variables explicatives.

La Figure 7 indique l'importance des variables, toutes stations et toutes heures confondues. On voit que le rayonnement Safran à la fin de l'heure sur laquelle porte le cumul et celui au début de cette heure (il s'agit de flux; les données horaires observées sont des cumuls) sont des variables très importantes. Une distorsion par rapport aux résultats précédents peut être liée à 3 causes :

- le nombre de stations sur lesquelles porte l'apprentissage a triplé
- la liste des prédicteurs est légèrement différente, par ajout de la durée d'insolation, disponible en presque toutes les stations à observation humaine, sauf Orly et Roissy.
- une évolution logicielle (simplification du décodage des fichiers de rayonnement Safran) dont l'influence est peu probable.

La durée d'insolation reste un prédicteur essentiel. Le rayonnement maximum théorique n'a pas d'influence à cause de la découpe en hauteurs de soleil. Par contre, l'heure semble avoir une importance trop grande. La visibilité, employée comme indicateur de présence d'aérosols, est beaucoup moins utile. L'influence non négligeable de la pluie peut être liée, soit à un réel phénomène physique, soit à un mauvais comportement du capteur en présence de pluie si les sels dessicants n'ont pas été renouvelés. En tout état de cause, il faut tenir compte de la pluie lors de l'établissement de liaisons statistiques par arbres, ne fût-ce que pour pouvoir être sûr que, quand il fait sec, les liaisons statistiques n'ont pas été contaminées par

ce défaut instrumental mal maîtrisé (c'est un des avantages de logiciels d'arbres d'isoler les sources de défauts).

Le mécanisme de traitement des données manquantes par variables de substitution a été activé pour Nantes, Roissy, Creil et Orly où la durée d'insolation n'est pas mesurée, ou rarement. Ces stations présentaient des scores de Nash très bas, liés au fait que ce mécanisme ne remédie pas de façon optimale à l'absence persistante d'une variable explicative, et peut-être aussi au fait que 3 de ces 4 stations n'ont pas de classement Radome (ou un mauvais). Sinon, si on fait aussi abstraction de Ajaccio, Bastia et Pau (ces deux dernières sont mal classées, le cas d'Ajaccio est difficile à comprendre), les coefficients de Nash dépassent 0.8 quasi systématiquement. Le bon comportement est particulièrement bien marqué à 12 h, où le rayonnement Safran ne fait pas l'objet d'une interpolation. Les biais sont aussi plus marqués à 0900 et 1500 UTC. La station la plus haute (Millau, 715 m) tout en disposant de séries de prédicteurs complètes ne présente pas de défaut ou de biais marqué (coefficients de Nash supérieurs à 93%).
Ce bon comportement à 1200 UTC se produit aussi avec des séries privées du rayonnement Safran, qui ne présentent pas de dégradation si la durée d'insolation est disponible.

4.3 Modes dégradés : jeu restreint de variables explicatives.

Il est très difficile, voire impossible, d'explorer toutes les façons de supprimer les 15 variables explicatives utilisées, aussi s'est on limité à 2 cas :

- le rayonnement Safran est absent, et on en a tenu compte lors de l'identification: pour les stations classiques, mesurant la durée d'insolation, le comportement reste très voisin. Les stations ne mesurant pas la durée d'insolation sont très dégradées. Ceci conforte le rôle primordial du rayonnement Safran, mais peut-être le traitement par arbres de régression n'est-il pas le meilleur. La durée d'insolation, si elle est présente, permet de s'affranchir de la disponibilité du rayonnement Safran.
- le rayonnement Safran et la durée d'insolation sont absents, et on en a tenu compte lors de l'identification : la dégradation a été jugée inacceptable.

A noter que faire fonctionner le mécanisme de variables de substitution ne mène pas au meilleur résultat : il faut disposer d'arbres identifiés pour cette configuration particulière qui, si ces modes dégradés étaient intéressants (pas dans le but de combler des zones géographiques vides de mesure), serait très lourd à traiter informatiquement.

Une autre façon d'aboutir à des résultats dégradés consiste à diminuer le nombre de stations servant à l'identification : une découpe de la France en 2 moitiés, l'une de climat océanique (le Nord Ouest), l'autre de climats plus continentaux, de montagne ou méditerranéen ne diminuait pas les biais, ce qui était contraire à ce que l'on espérait en travaillant dans des zones plus homogènes, mais augmentait très légèrement la variance. Noter enfin que les deux stations dont la suppression a l'impact le moins négatif sont, dans l'ordre, Bastia et Ajaccio.

Chapitre 5

Krigeage

On a vu précédemment que le rayonnement reconstitué par Safran est très utile, conjointement à des observations humaines, pour reconstituer du rayonnement visible en la majorité des stations de France continentale. Cette utilité est tempérée par le fait que, si on utilise aussi la durée d'insolation, la suppression du rayonnement Safran n'induit pas de conséquences significatives. On peut donc disposer alors de rayonnement reconstitué dans toutes les stations à observation humaine mesurant la durée d'insolation (une par département), ce qui n'est pas négligeable devant le nombre de stations Radome de qualité appréciable mesurant ce rayonnement[8]. Par contre, le problème d'affecter à chaque maille Safran (des carrés de 8 km de coté) une valeur de rayonnement peut être résolu partiellement par krigeage, au moins dans les zones de plaine, où il y a suffisamment de mesures pour une validation quelque peu convaincante.

5.1 Généralités

On a choisi des méthodes très classiques de krigeage ordinaire (respectant la moyenne) pour des raisons de disponibilité et de simplicité. Les particularités les plus notables sont les suivantes :

- Du fait du grand nombre de données, la solution consistant à identifier la forme d'un variogramme pour chaque heure a été exclue. La possibilité existe dans ***gstat***, mais cette identification fonctionne de façon réaliste dans 95 % des cas, le reste -qui n'est pas du tout négligeable sur des données

[8] à terme, 2 par département; le choix du réseau Radome réside dans la maîtrise des conditions d'implantation et de maintenance, ainsi que de la possibilité de disposer des observations en temps réel

horaires- imposant un contrôle subjectif très lourd.

- Les variogrammes adimensionnés (ce point implique que l'on ne s'intéresse pas aux unités des carrés des différences et ne serait gênant que si on souhaitait estimer une variance de reconstitution) sont identifiés sur des carrés de différences de rayonnements horaires, regroupés dans des classes de distances, voire de directions et cumulés sur un mois (à 1200 UTC, pour simplifier). C'est une structure très proche de celle retenue pour l'action J98b de spatialisation des pluies.
- Au vu d'une dizaine de variogrammes adimensionnés mensuels, on a pu constater que les variogrammes se stabilisaient fréquemment à l'infini surtout après prise en compte d'un terme peu physique de dérive avec les positions et que l'anisotropie[9] était faible; on pouvait modéliser ce variogramme, à une constante multiplicative près (peu intéressante sauf si on voulait tracer des cartes de variance modélisée par krigeage) sous la forme d'un effet de pépite, qui représente environ 20% de la variance et d'un terme exponentiel de pente à l'origine 1/110 km[10].
- On a augmenté un peu arbitrairement la sensibilité en fixant l'effet de pépite à 13% dans tout ce qui suit (ce qui permet de mieux respecter les valeurs ponctuelles).
- On a aussi négligé le terme de dérive (comme combinaison linéaire variant avec le jour des coordonnées géographiques, qui peut se manifester si, par exemple, le temps est très ensoleillé en Bretagne et très couvert en Provence) en ne travaillant que dans des zones glissantes de largeur 300 km et sur au maximum 15 postes (c'est aussi une optimisation du temps de calcul d'un facteur 100; il est par ailleurs difficile de comprendre pourquoi le rayonnement en Alsace serait influencé par celui en Bretagne et en Provence); c'est ce qu'on appelle un krigeage local.

Les logiciels employés ont été (successivement, en parallèle, puis seulement ***gslib***)

- ***gstat*** (Pebesma 2003), un greffon de ***R*** qui est le seul (parmi 3 autres logiciels de krigeage disponibles avec ***R*** en 2004, dont un raisonnablement débuggé) à offrir cette possibilité de krigeage local.
- ***gslib*** (Deutsch & Journel 1992; cours et mode d'emploi en ligne dans Landim et Ribeiro 2002), très souvent utilisé et cité, déjà utilisé à Météo-France (action J98b du Retic) et dont un des modules (***kt3b***) offre des possibilités de validation suffisamment développées pour pouvoir comparer ses propres sorties avec celles d'autres logiciels (dans le cas présent, où les équations à résoudre étaient les mêmes, on a ainsi pu vérifier que ***gstat*** et

[9] Anisotropie = variation de la forme du variogramme avec la direction.

[10] un variogramme de la forme exp(-d/p) a une pente à l'origine de 1/p, p étant la portée (dont l'unité est ici fixée en km pour pouvoir se comparer si besoin à la littérature).

gslib donnaient les mêmes résultats). Une interface minimaliste avec ***R*** (et avec PVwave) a été écrite.

A titre d'exemple, on a affiché un variogramme mensuel pour le mois d'août 1998 à 1200 UTC (Figure 8). L'axe horizontal est la distance (en hm), l'axe vertical est le cumul des carrés des différences de rayonnement.

On présentera les cartes de rayonnement spatialisé le 1 octobre 1997, dans 3 cas de figure :

- les points d'appui du krigeage sont limités aux postes Radome de classe inférieure ou égale à 3.
- on complète ces points d'appui par des reconstitutions systématiques par arbres (i.e. les points disposant à la fois de mesures de rayonnement et de possibilités de reconstruction voient leur observation de rayonnement reconstituée d'office; cette façon de faire est caricaturale et ne prétend pas faire l'objet d'une application en routine, mais elle permet de voir d'éventuels défauts).
- Safran, pour information.

5.2 Situation du 1 octobre 1997

Il s'agit d'une situation avec un ensoleillement exceptionnel sur la majeure partie de la France, à l'exception de quelques zones : Nord-Pas de Calais et Ardennes (nuages bas), côtes du Pays de Caux et de Picardie (brouillards épais tardant à se dissiper). Au Nord de la Seine et jusqu'en Alsace, des nuages moyens peuvent voiler le ciel.

La Figure 9, où un réseau lâche sert de point d'appui au krigeage, montre que les entrées d'air maritimes sur le Pays de Caux ne sont pas discernées avec ce réseau lâche.

On a indiqué en noir les observations servant de point d'appui au krigeage, en rouge celles dont la qualité des sites ou des particularités de gestion font qu'elles ne sont pas jugées dignes d'être utilisées pour une spatialisation (c'est discutable à 1200 UTC; cependant, elles risquent de ne pas être transmises en temps réel) mais qui ne sont pas moins utiles pour une vérification. Les points notés 'x' disposent d'observations humaines et de durée d'insolation et serviront de nouveaux points d'appuis au krigeage dans la carte suivante. Ils sont particulièrement nombreux et visibles en Picardie, Beauce et Pays de Caux, régions où il n'y a pas de mesures de

rayonnement.

La carte de la Figure 9 diffère beaucoup de celle de la Figure 10 dans le pays de Caux (cette dernière semble plus réaliste en ce qui concerne les entrées d'air maritime près du Nord de la Manche, ceci au vu du BHER et de la carte Safran qui va suivre). Noter aussi que les estimations par arbres des rayonnements ponctuels sont très voisines de la réalité à Perpignan, Carcassonne, Rennes, Brest Reims et La Rochelle, ceci serait lié à la simplicité des couches nuageuses.

La carte de la Figure 11 montre les rayonnements Safran correspondant : les limites sont beaucoup plus nettes, les maxima le long des sommets des Pyrénées semblent correspondre à une réalité physique (sans que nous ayons de mesures pour le confirmer). Le minimum correspondant aux entrées d'air maritimes près de la Manche semble réaliste. Par contre, le minimum au Sud du Vercors et celui près de la Camargue sont très douteux (la mesure à Montpellier confirmerait le caractère douteux de ce dernier minimum).

Chapitre 6

Validation sur une période (2004) et un réseau n'ayant servi à aucun apprentissage

6.1 Généralités

Quatre méthodes de reconstitution du rayonnement horaire incident ont été testées sur 2004, année qui sert de vrai fichier de validation, avec onze stations ne rentrant pas dans la BDCLIM. On dispose ainsi d'un échantillon indépendant, tant par sa date que par le fait que ces stations ne sont pas soumises à une vérification par cohérence spatiale, ce qui peut favoriser des techniques de krigeage. Ces méthodes sont les suivantes :

- interpoler spatialement par krigeage les stations Radome de niveau inférieur ou égal à 3, jugées de qualité acceptable (comme pour la carte 9 du Chapitre 5).

- "arbres systématiques" : ajout au réseau de stations de mesure de rayonnement des points fictifs, où on dispose de mesures de durée d'insolation et de nature et étendue des couches nuageuses; le rayonnement y est reconstitué par des ensembles d'arbres de régression identifiés sur la période 1995-2003, en écrasant toute valeur du rayonnement solaire incident préexistante (comme pour la carte 10 du Chapitre 5).

- la troisième méthode est très proche de la seconde, mais moins caricaturale et a priori plus proche de choix opérationnels : si on dispose de mesures de rayonnement en une station de niveau Radome inférieur ou égal à 3, on la garde (les stations de Nice -masquage par montagne- et Roissy -peupliers- sont de niveau 4 et font donc l'objet de reconstitutions).

- la dernière, mise en annexe, consiste à comparer les observations indépendantes avec des sorties des flux horaires de Safran.

Les utilitaires de krigeage fabriquent, avec une résolution de 8 km (maille Safran avant stockage dans la BDAP), des champs de rayonnement solaire incident cumulé sur l'heure. Les points les plus proches des nouveaux postes sont ensuite extraits de ces cartes, à des fins de comparaison. Ces points sont matérialisés informatiquement par des fichiers datés et géoréférencés.

Les critères de dépouillement affichés sont les suivants, dans tout ce qui suit :

- affichage par biplot (diagramme de dispersion).

- tracé de séries temporelles sur 4 semaines, une par saison.

- affichage des biais, écart-quadratiques (EQM), coefficients de corrélation et critères de Nash (% de variance expliquée, si les biais sont faibles, complément à 1 du ratio entre l'EQM et la variance de la grandeur à prévoir dans le cas général); ces indicateurs de qualité sont collationnés/synthétisés dans des tables html.

A ce propos, les calculs sont faits pour des paires (données réelles/données observées) non manquantes, avec l'actuelle convention de la BDCLIM, où les données d'irradiation solaire de nuit sont systématiquement manquantes. Le recalcul des biais (resp. écarts quadratiques), si on voulait tenir compte de ces cas triviaux, s'obtiendrait par une simple division par deux (resp. 1.4) de ce qui est affiché sur une base de calculs toutes heures diurnes et saisons confondues (ces cas de rayonnement de nuit peuvent être pris en compte dans les calculs de bilans, ceci peut gêner des intercomparaisons de méthodes; on a choisi de respecter la présence dans la BDCLIM comme base de calcul).

Malgré cela, on sera peut être surpris par la valeur insolemment élevée des coefficients de corrélation et critères de Nash : elle découle de la présence de cycles diurnes et annuels. Un dépouillement saison par saison et heure par heure serait plus vertueux, mais plus difficile à interpréter et plus fastidieux.

6.2 Position des stations de validation

La carte de la Figure 12 indique les noms et emplacements des postes utilisés, par rapport au réseau de mesures de Météo-France. On distingue les postes Radome de code inférieur ou égal à trois, servant de points d'appui au krigeage, les autres points de mesure de rayonnement, qui ne sont pas transmis en temps réel ou dont le site est entaché de masques (code Radome 4 ou absent), et les points pouvant faire

l'objet d'une reconstitution, éventuellement confondus avec les points de mesure (ne serait-ce que pour identifier des relations).

On constate que les stations de vérification (hors BDCLIM) sont essentiellement implantées dans le Sud de la France, et au voisinage de réseaux de mesure assez denses[11] (sauf Luxey). Noter aussi que ces stations sont situées en dessous de 500 m d'altitude (à l'exception de Sault).

6.3 Dépouillement

Trois essais de spatialisation sont décrits, par ordre de complexité croissante. Le caractère indiscernable des 3 algorithmes de reconstitution y est mis en évidence. Une visualisation (forcément sommaire) des résultats est donnée par la Figure 13 (tracé des séries temporelles observées et reconstituées).

6.3.1 Krigeage sans reconstitution.

A titre d'exemple, un faible déphasage apparaît à Smosrex; il n'est pas visible sur les séries temporelles, mais soupçonné au vu des biplots (séparation par coloriage entre réponses le matin et l'après midi) et découle vraisemblablement d'un malentendu sur les temps d'intégration (passage de flux instantanés à des cumuls). Le fait qu'on l'ait détecté tardivement interdit de le corriger. Une tendance à sous-estimer les variations d'une heure à l'autre se manifeste par ciel couvert (par exemple le 7 septembre ou le 25 novembre); elle est attribuée au lissage induit par le krigeage (somme pondérée des valeurs des postes voisins, les poids se sommant à 1) ; cela peut susciter des réserves lors, par exemple, d'averses localisées, où le réseau d'observations utilisé est trop lâche.

En ce qui concerne les particularités des observations, Smosrex (au moins le 21 juin) a un comportement sain au voisinage des données manquantes, ce qui le différencie de Gif/Yvette au mois d'avril, où des valeurs anormalement élevées suivent et précèdent des valeurs manquantes, ce qui indiquerait un degré de sophistication élevé de Smosrex. Le faible nombre de valeurs jugées aberrantes n'incite pas à l'usage d'autres scores (EAM) ni à rejeter cette station.

[11] La forte densité de points de mesure au voisinage de Hesse s'explique par le fait que la table des stations Radome n'était pas à jour pour la region NE : plutôt que de censurer cette région, on a préféré décider que toutes les stations de cette région étaient de très bonne qualité.

STATION	corr	biais	Nash	EQM
SMOSREX	0.96	0.46	0.92	26.73
Hesse	0.97	-4.37	0.94	20.94
Salon	0.98	4.54	0.96	20.91
Saint-Gilles	0.96	-8.85	0.91	33.51
Marseillan	0.95	-6.94	0.89	34.69
Roujan	0.94	-14.80	0.84	44.46
Luxey	0.96	-2.06	0.93	24.92
Creysse	0.96	-2.51	0.91	29.12
Sault	0.94	-4.17	0.89	35.71
Gif	0.91	1.66	0.82	36.53
Puechabon	0.90	0.80	0.82	43.64

TABLEAU 3 - Quelques indicateurs de performance (Krigeage seul).

Il n'y a pas de déphasage visible à Hesse; la très bonne qualité des reconstitutions en ce point peut s'expliquer parce que cette station a été utilisée pour des campagnes de mesure récentes, et par le fait que l'on soit dans un réseau artificiellement dense. En effet, n'ayant pas accès en 2004 aux informations concernant la qualité et le label des stations servant de point d'appui au krigeage, on a admis que toutes les stations étaient de bonne qualité[12]. La variabilité par ciel changeant est aussi très atténuée (ex. 18 septembre) en cette station.

Les autres stations offrent des reconstitutions de qualité inégale et intermédiaire entre celle de Hesse et celle de Gif.

[12] Ceci impliquerait que les niveaux Radome sont trop stricts pour reconstituer une station, la généralisation de cette affirmation à l'échelle de la France demanderait quelque courage : affirmer qu'une erreur accidentelle dans la qualification des stations serait tolérable, au vu des autres causes d'erreurs, est peut être plus acceptable.

6.3.2 Comparaison entre réalité et reconstitution systématique par arbres, suivie de krigeage.

Il est difficile de mettre en évidence une différence avec la méthode précédente au simple vu des séries temporelles; cependant, les scores sont généralement un peu moins bons que précédemment, ce qui peut être imputé à 2 causes :

- les postes sont situés dans des zones d'observations plus denses que la Haute-Normandie ou la Beauce.
- le choix des arbres, construits uniquement avec des données d'observation, induit forcément une dégradation.

La dégradation, faible, se manifeste plutôt par ciel couvert (ex. le 16 avril à Marseillan).

La diminution du biais à Hesse est peut être liée au forçage à un réseau pseudo-radome dense de stations moins homogènes, faute d'information officielle (en 2004).

Cette interprétation est mise en cause par la diminution des biais dans le Languedoc, à Sault et à Luxey, au prix d'une petite augmentation de l'EQM partout.

A signaler aussi que cette méthode, quoique caricaturale, avait été trouvée légèrement supérieure à la méthode présentée au paragraphe 6.3.1 (Krigeage sans reconstitutions) sur les séries de la BDCLIM entre 1997 et 2003, ce tant pour les postes non Radome que pour les postes Radome (on veillait alors à ce que, provisoirement, ils ne servent pas de points d'appui au krigeage). Cette contradiction est imputable à 3 causes, pas forcément exclusives :

- On s'était limité aux heures synoptiques (0900, 1200 et 1500 UTC), pour davantage de stations.
- On disposait d'un réseau plus irrégulier et moins dense.
- On peut aussi envisager un passage involontaire sur des périodes ayant servi à caler des arbres, d'où l'introduction d'un biais d'optimisme.

6.3.3 Comparaison entre réalité et reconstitution par arbres limitée, suivie de krigeage.

Il s'agit d'un compromis entre les 2 premières méthodes et, à ce titre, les résultats, sont encore plus indiscernables que précédemment.

Au vu des tableaux de résumés (la reconstitution en question) et du krigeage simple, elle se traduit par :

- une amélioration de la reconstitution à Smosrex et Puechabon tous scores

confondus (sauf un léger biais)

- une diminution des biais à Salon, St Gilles, Marseillan, Roujan, Luxey et Sault, soit plus de la moitié des postes. La diminution du biais et de l'EQM à Hesse est parfaitement négligeable, du fait du choix d'un réseau artificiellement dense dans cette région.

STATION	corr	biais	Nash	EQM
SMOSREX	0.96	*0.65*	0.92	**25.61**
Hesse	0.97	**-4.32**	0.94	**20.92**
Salon	0.98	*5.21*	0.96	*21.32*
Saint-Gilles	0.95	**-4.18**	0.90	*35.10*
Marseillan	0.95	**-3.68**	0.89	*35.55*
Roujan	0.93	**-11.30**	0.84	*44.75*
Luxey	0.95	**0.69**	0.90	*29.42*
Creysse	0.95	*-3.88*	0.91	*29.65*
Sault	0.94	**-3.29**	0.89	**35.70**
Gif	0.89	*3.97*	0.78	*40.84*
Puechabon	0.93	*4.72*	0.87	**37.38**

TABLEAU 4 - Reconstitution prudente par arbre puis krigeage (on a mis en italique les indicateurs dégradés, même très peu, par rapport à un krigeage sans reconstitution, en gras ceux qui sont améliorés).

6.4 Conclusions

On a pu vérifier que le krigeage permet de reconstituer avec réalisme des séries horaires de rayonnement hors BDCLIM (il est par ailleurs difficile de valider de façon convaincante un krigeage dans la BDCLIM où les données sont vérifiées par cohérence spatiale avant stockage). L'apport des reconstitutions par arbres est très faible en valeur absolue (sans changer la forme du krigeage : tenir compte du fait que ces reconstitutions sont forcément bruitées impliquerait l'emploi de co-krigeage, peu utilisé et que l'on a omis par souci de simplicité), et les stations utilisées sont situées dans des régions denses (ce n'est pas le cas de la

Beauce...) et ne permet pas de conclure de façon définitive, sauf à une diminution des biais jointe à une légère augmentation du bruit. En tous cas, il n'induit pas de dégradation visible. Le cas des régions en altitude (Hautes Alpes) doit naturellement être exclu de ces conclusions. Par ailleurs, quelle que soit la méthode utilisée, il semblerait qu'elle surpasse, du moins en 2004 en plaine, des sorties non corrigées de Safran (voir annexe).

Chapitre 7

Conclusion

Les observations humaines et de durée d'insolation permettent de compléter un réseau d'observations de rayonnement global initialement assez lâche et irrégulier, pour servir de points d'appui à un krigeage simple sans introduire de dégradation notable, même dans le pire des cas. Les améliorations potentielles sont les suivantes (ordre arbitraire et sans prétention à l'exhaustivité) :

- incorporation de données prévues dans les reconstitutions ponctuelles, par arbres (ou par tout autre méthode statistique si besoin) : il s'agit de données très facilement accessibles en temps réel.
- prise en compte du relief :Safran en tient compte, alors que les réseaux d'observation servant à l'identification et à la validation sont situés plutôt en plaine.
- emploi de co-krigeage à la place du krigeage pour séparer en bonne logique les points reconstitués des mesures réelles.

Remerciements

A. Granier (UMR-INRA UHP 1137) a fourni les données de Hesse, J.M. Ourcival (CEFE-CNRS UMR 5175) celles de Puechabon , les données des autres stations de validation ont été données par F. Huard (INRA AgroClim).
L'emploi du krigeage a été préconisé par G.Therry(Retic/D).
Le fait que le filtrage subjectif des données induise une dégradation des performances lors de l'apprentissage (par perte d'information) et soit très lourd a été suggéré par G. Oppenheim (CNRS/Orsay).

Références

Références bibliographiques

Bankert,R.L., Hadjimichael, M., Kuciauskas, A.P., Thomson, W.T. & Richardson, K. 2004. Remote cloud ceiling assessment using Data Mining methods JAM dec 2004 V43.12 pp 1929-1945.

Bel, L., Bellanger, L., Bonneau, V., Ciuperca, G., Dacunha-Castelle, D., Deniau, C., Ghattas, B., Misiti, M., Misiti, Y., Oppenheim, G., Poggi, J-M., Tomassone, R., 1999. Eléments de comparaison de prévisions statistiques des pics d'ozone. RSA v XLVII(3) 7-25.

Black, J. N., Bonython, C. W., Prescott, J. A., 1954. Solar radiation and the duration of sunshine. QJRMS pp 231-235 ,1954.

Boone, A., Habets, F., Noilhan, J., Blyth, E., Dirmeyer, P., Gusev, Y., Haddeland, L., Koster, R., Lohmann,D., Mahanama, S., Mitchell, K., Nasonova, O., Niu, G. Y., Pitman, A., Polcher, J., Shmakin, A. B., Tanaka, K., van den Hurk, B., Verant, S., Verseghy, D. & Viterbo, P., 2002. The Rhône-Aggregation Land Surface Scheme Intercomparaison Project. JoC, 2002.

Breiman L., Friedman J.H., Olshen R. & Stone C.J., 1984, Classification and Regression Trees . Wadworth, Belmont CA.

Breiman, L., 1994 . Bagging Predictors; Tech. Rep. 421, Berkeley, Dept. of Statistics.

Buhlman, P. & Yu, 2001, B. : Analysing bagging. Annals of Statistics 30, pp 927-961

Burrows W.,1991. Objective guidance fo 0-24h & 24-48h mesoscale forecasts of lake effect snow using CART. Weather & Forecasting, 1991, V9, 357-378.

Canellas C., Merlier C. et Perarnaud V. (1994) Le gisement solaire en France: Estimation de l'irradiation solaire globale reçue sur une surface horizontale. Note SCEM.

der Megreditchian, G..1993 Le traitement statistique des données multidimensionnelles Tome 2 p 82 Direction de la Météorologie Nationale.

Deutsch, CV, Journel, AG. GSLIB: geostatistical software library and users guide
OUP 1992.

Durand, Y., Brun, E., Merindol, L., Guyomarch, G., Lesaffre, B., Martin, E., 1993 : A meteorological estimation of relevant parameters for snow models. Annals of Glaciology, 18: pp 65-71, 1993.

Etchevers, P. Modèlisation de la phase continentale du cycle de l'eau àl'échelle régionale. Impact de la modèlisation de la neige sur l'hydrologie du Rhône. Master's thesis, Université P.Sabatier de Toulouse, 2000.

Ghattas, B. 1999 Prévision des pics d'ozone par arbres de régression simples et agrégés par bootstrap. RSA, v XLVII 61-80

Ghattas, B. 1999 Importance des variables dans les méthodes CART. Modulad 24,pp 29-39.

Hastie,T., Tibshirani, R.J & Friedman,J.H. 2001 The elements of statistical learning : data mining, inference and prediction. Springer Series of Statistics. NY Springer 549pp.

Hunt, L.A., Kuchar, L. & Swanton, C.J., 1998. Estimation of solar radiation for use in crop modelling. Agr.&For. Meteorology. 91(1998) 293-300

Golaz-Cavazzi, C. Modèlisation hydrologique à l'échelle régionale appliquée au bassin du Rhône. Master's thesis, Ecole Nationale Supérieure des Mines de Paris, 1999.

Ihaka & Gentleman, 1996 R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics 5: 299-314.

Ledoux, E., Girard, G., de Marsilly, G., Deschene, J . Spatially distributed modelling: conceptual approach, coupling surface water and grounwater. Unsaturated flow hydrologic modeling theory and practice. NATO, ASI Series C, 275:435-454,1989.

Le Moigne, P. : Description de l'analyse des champs de surface sur la France par le Système Safran. Note GMME no 77, dec.2002.

Loh, W. Y. Regression Trees with unbiased variable selection and Interaction Detection . Statistica Sinica 2002, v.12 pp. 361-386

Morel, S. Modélisation distribuée du bilan hydrique à l'échelle régionale: application au bassin Adour Garonne. Master's thesis, Université P. Sabatier de Toulouse, 2002

Noilhan, J. & Planton, S.: A simple parametrisation of land surface processes for météorological models. MWR, 117:536-549,1989.

OMM, 1987 : Aspects météorologiques de l'utilisation du rayonnement solaire comme source d'énergie. OMM No 557, NT 172,1997.

Ritter, B. & Geleyn, J.-F., 1992 . A comprehensive radiation scheme for numerical weather prediction models with potential applications in climate simulations. MWR, 120-2: pp 303-323.

Ryan W.F., 1995. Forecasting severe ozone episodes in the Baltimore metropolitan area. Atm. Env. 29(17) 2387-2398.

Soler , A. 1990.Statistical comparison for 77 european stations of 7 sunshine based models.
Solar energy, V41,No 6 pp 363-370, 1990

Stone C.J., Hansen, M. H., Kooperberg, C. & Truong, Y. K., 1997 Polynomial splines and their tensor products in extensed linear modeling (avec remarques) 1994 Wald Memorial Lecture, The Annals of Statistics v.25, n.4 pp 1371-1490.

Zorita, E. & von Storch, H. , 1999. The analog Method as a simple statistical downscaling technique: comparison with more complicated methods. JOC 1999/08 v.12 pp2474-2489.

Autres références

- BHER Bulletin hebdomadaire d'études et de renseignements.
- Landim PMB, Monteiro RC. Introduçao ao GSLIB
www.rc.unesp.br/gce/aplicada/gslib.pdf
- R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
- Rpart: Terry M Therneau and Beth Atkinson. R port by Brian Ripley <ripley@stats.ox.ac.uk>. (2005). rpart: Recursive Partitioning. R package version 3.1-22. S-PLUS 6.x original at http://www.mayo.edu/hsr/Sfunc.html

Liste des figures

ANNEXE 1

Critères de choix informatique

Il y a des quantités de logiciels d'identification disponibles dans le domaine public et une mise en concurrence systématique serait très lourde; aussi a-t-on fait le tri parmi une vingtaine de logiciels (en 2003; actuellement, ils sont près de 200...) avec les critères de sélection suivants, l'ordre d'énumération n'étant pas celui des priorités éventuelles :

- être installable sur le maximum de calculateurs modernes
- être utilisé par ailleurs à Météo-France, ou être de grande diffusion.
- pouvoir fournir des éléments d'interprétation; en effet, la qualité des intrants étant mal connue, il est rassurant de vérifier si une équation de régression, par exemple, est loin d'être absurde. Ceci exclut les méthodes des analogues et les réseaux de neurones.
- être dépourvu d'erreurs informatiques gênantes.
- ne pas faire appel à des générateurs de nombres aléatoires, du fait de mauvais souvenirs liés à une absence d'initialisation rendant les résultats peu reproductibles. Cette critique peut être tempérée par le fait que tous les exemples d'utilisation de logiciels modernes rappellent comment initialiser ces tirages aléatoires. Ceci exclut les stabilisations par 'bagging' et les réseaux de neurones.
- ne pas avoir de mauvais comportement en présence d'éventuelles valeurs aberrantes, si elles restent rares.
- gérer aussi bien des grandeurs continues que des codes, caractérisant par exemple le type de nuage (un ciel entièrement couvert par des nuages ténus laisse passer plus de rayonnement qu'un ciel obscurci par des nuages épais). Ceci défavorise les réseaux de neurones, les méthodes des analogues (il est difficile de définir objectivement une distance entre variables explicatives

codées) et les anciens logiciels de régressions linéaires.

- offrir des temps d'identification raisonnables. Ceci défavorise les réseaux de neurones (il peut être nécessaire de les réinitialiser plusieurs dizaines de fois pour aboutir à un optimum pas trop local...) et la fixation de méta-paramètres par optimisation systématique.

ANNEXE 2

Comparaison de Safran avec les observations de 11 postes hors BDCLIM en 2004

On a comparé de la même façon qu'en 6.3 les observations de flux en nos 11 stations aux estimations de flux (et non de cumuls) données par Safran, le seul traitement autre qu'un décodage compliqué de la BDAP consistait à un changement d'unité, à des fins de comparaison confortable, et de signe (les valeurs de la BDAP ont un sens physique). Deux types de comparaison ont été faits, l'un avec une correction de datation d'une ½ heure pour Smosrex, l'autre sans (on ne pouvait pas le faire dans la section 6 sans introduire un élément de subjectivité peut-être choquant).

On peut remarquer un gain de qualité certain mais faible lorsque l'on veille à une datation correcte (sauf pour le biais: le contraire serait inquiétant, en indiquant qu'un filtrage simple ne respecterait pas la moyenne), mais les scores restent aussi très inférieurs à ce que l'on obtient en section 6. Le tracé, Figure 14, des évolutions temporelles observées, modélisées par Safran et spatialisées après reconstitution montre que, dans la plupart des cas, les séries de rayonnement SAFRAN s'écartent davantage de l'observation que les reconstitutions statistiques. Ceci est imputé à la complexité du décodage de la BDAP jointe à la nouveauté des logiciels de validation pour le rayonnement descendant de Safran .

STATION	cor	biais	Nash	EQM
Saint-Gilles	0.90	-25.40	0.73	54.66
Smosrex	0.88	-10.50	0.77	44.35
Hesse	0.84	-12.00	0.69	48.57
Salon	0.91	-11.20	0.81	43.23
Luxey	0.85	-17.40	0.69	51.40
Roujan	0.84	-32.80	0.60	69.68
Sault	0.87	-27.60	0.68	59.37
Creysse	0.88	-16.90	0.73	49.70
Gif	0.81	-11.50	0.64	51.57
Puechabon	0.89	-14.50	0.76	49.36
Marseillan	0.87	-21.50	0.70	56.87

TABLEAU A1 - Scores Safran avant correction de datation. Reconstitution Safran.

STATION	cor	biais	Nash	EQM
Saint-Gilles	0.93	-24.10	0.79	49.92
Smosrex	0.90	-10.00	0.81	41.16
Hesse	0.88	-11.40	0.75	43.82
Salon	0.94	-10.80	0.86	37.32
Luxey	0.89	-16.60	0.76	45.92
Roujan	0.89	-31.00	0.68	63.11
Sault	0.83	-28.80	0.61	65.83
Creysse	0.91	-16.10	0.79	44.71
Gif	0.84	-11.20	0.69	48.06
Puechabon	0.89	-13.40	0.77	49.44
Marseillan	0.91	-20.60	0.77	50.67

TABLEAU A2 - Scores Safran après correction de datation. Reconstitution Safran avec filtrage temporel.

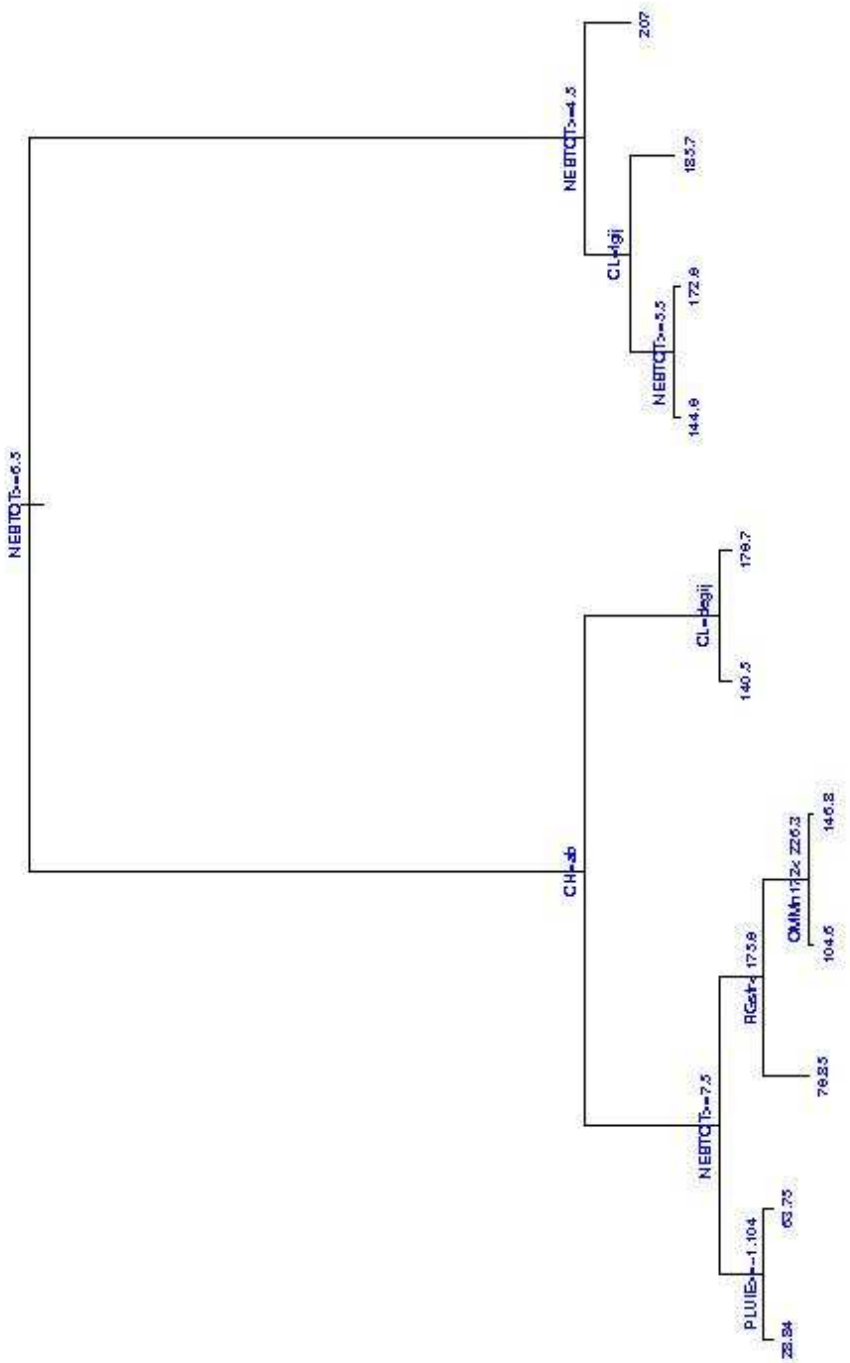
NEBTQT>=6.3
NEBTQT>=4.5
207
CL=lgi
183.7
NEBTQT>=5.5
172.9
144.9
CL=lgii
179.7
140.5
CH=ab
NEBTQT>=7.5
RGst=1
173.9
CMMn1=2
226.3
146.8
104.6
76.85
PLUIE>=1.104
63.75
28.84

FIGURE 1

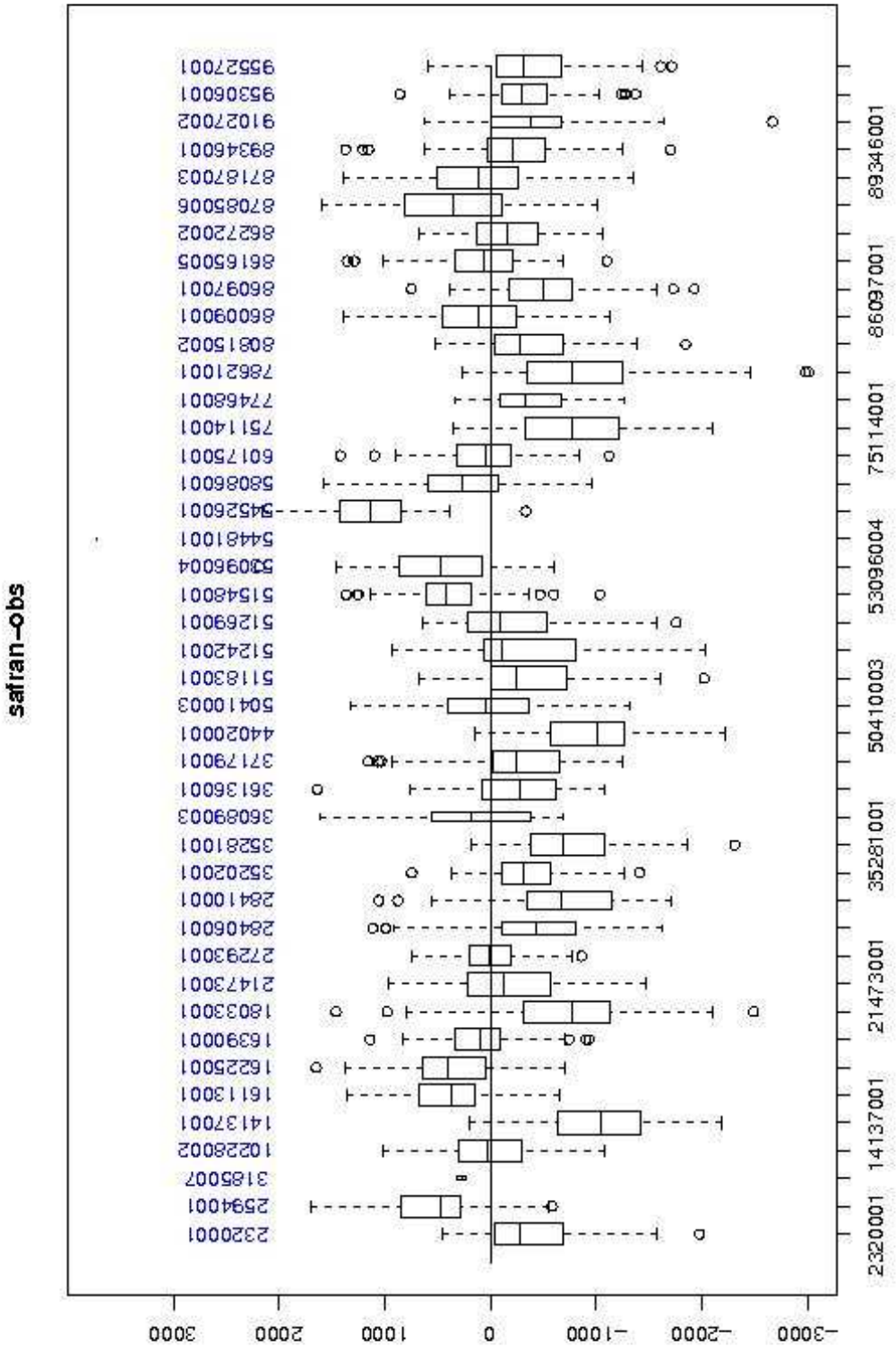
safran−obs
9552700 1
9530600 1
9102700 2
8934600 1
8718700 3
8708500 6
8627200 2
8616500 5
8609700 1
8600900 1
8081500 2
7862100 1
7746800 1
7511400 1
6017500 1
5808600 1
5452600 1
5448100 1
5309600 4
5154800 1
5126900 1
5124200 1
5118300 1
5041000 3
4402000 1
3717900 1
3613600 1
3608900 3
3528100 1
3520200 1
2841000 1
2840600 1
2729300 1
2147300 1
1803300 1
1639000 1
1622500 1
1611300 1
1413700 1
1022800 2
3185007
2594001
2320001
2320001
14137001
21473001
35281001
50410003
53096004
75114001
86097001
89346001
3000
2000
1000
0
−1000
−2000
−3000

FIGURE 2

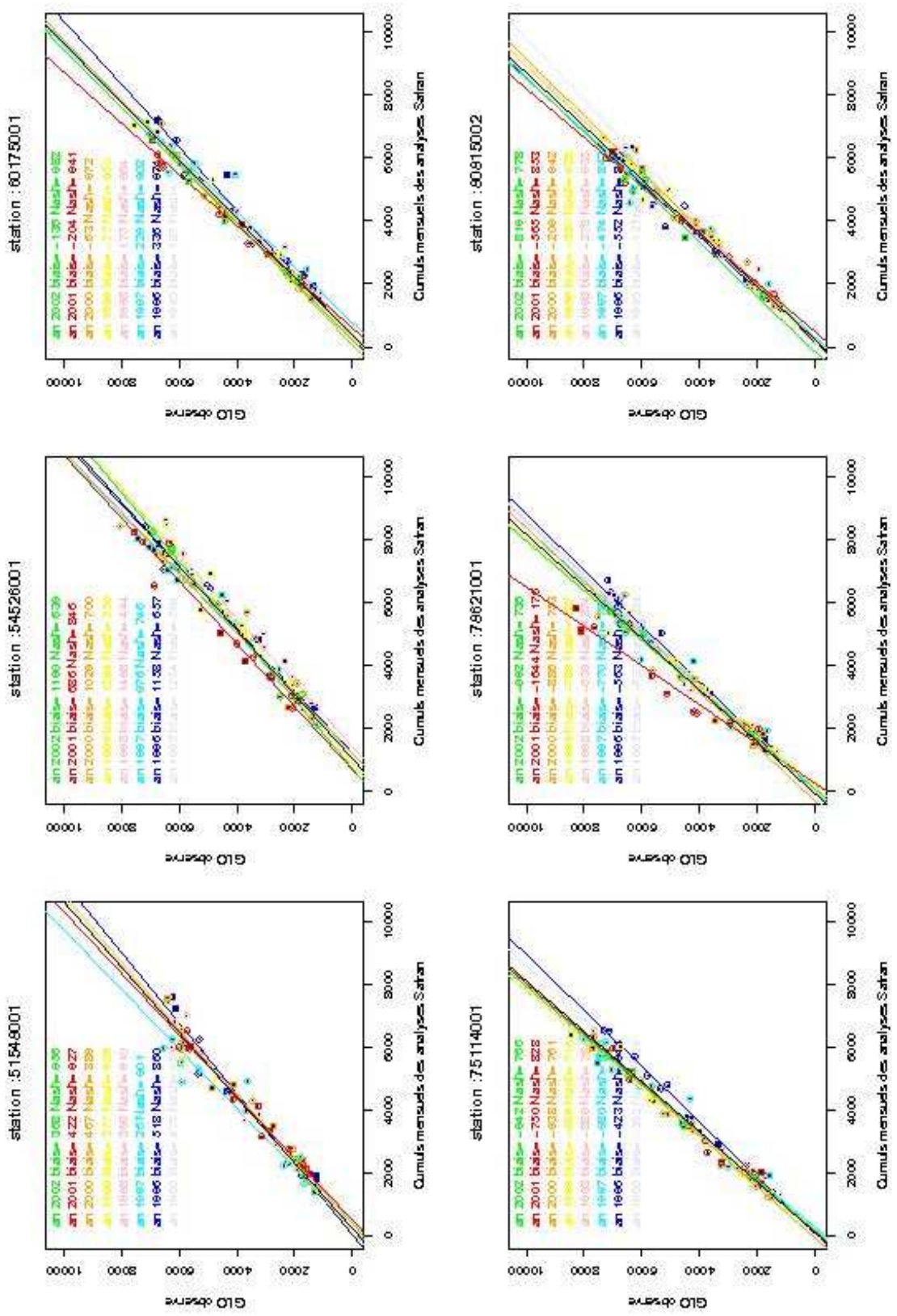
station : 80175001
station :54526001
station :51548001
station :90815002
station :78621001
station :75114001
Cumuls mensuels des analyses Safran
GIO observe

FIGURE 3

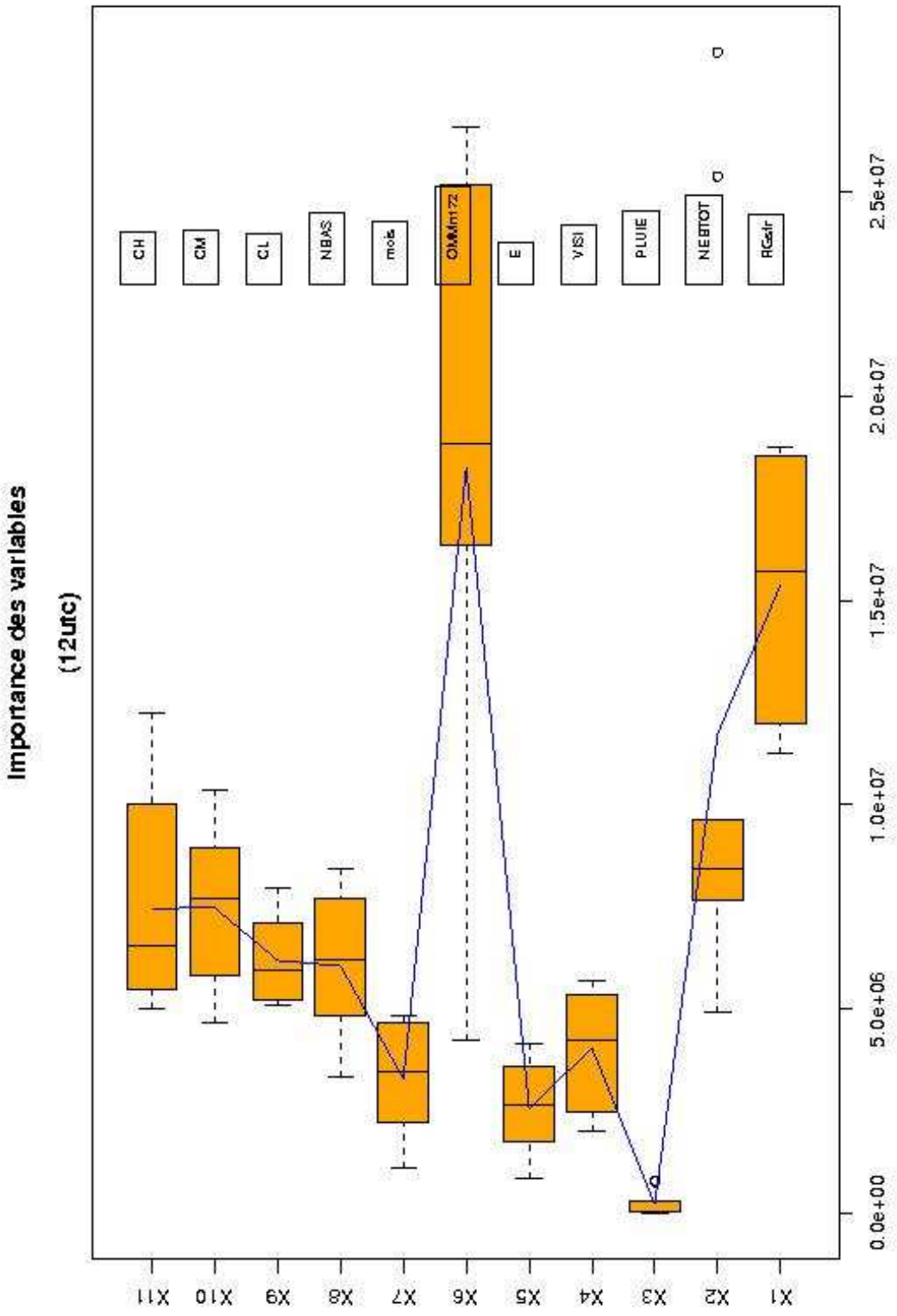
Importance des variables
(12utc)
CH
CM
CL
NBAS
mois
OLMdn172
E
VISI
PLUIE
NEBTOT
RGclr
X11
X10
X9
X8
X7
X6
X5
X4
X3
X2
X1
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07

FIGURE 4

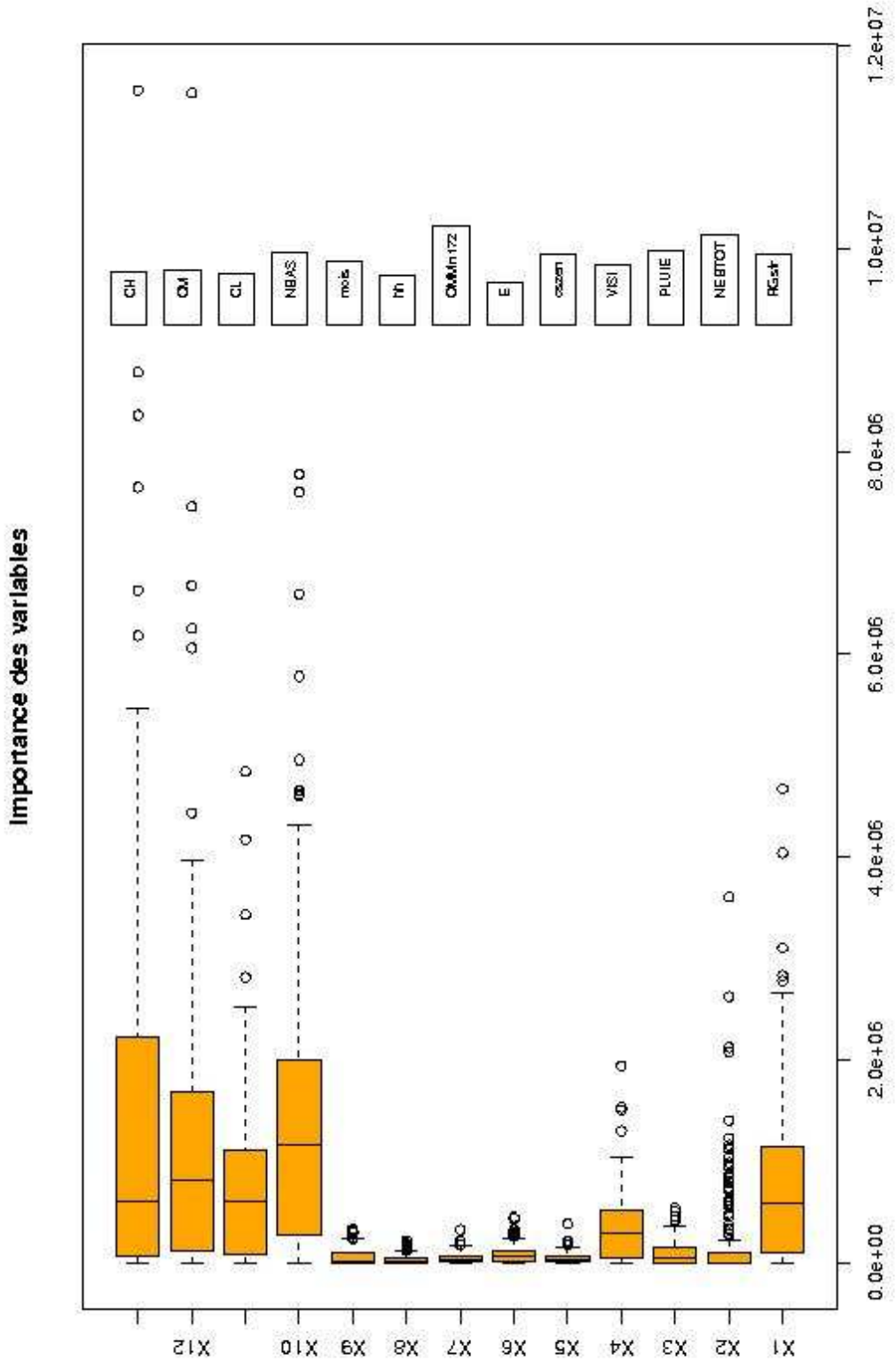
Importance des variables
CH
CM
CL
NBAS
mois
hh
DMMin172
E
casm
VISI
PLUIE
NEBTOT
RGsit
X12
X10
X9
X8
X7
X6
X5
X4
X3
X2
X1
0.0e+00
2.0e+06
4.0e+06
6.0e+06
8.0e+06
1.0e+07
1.2e+07

FIGURE 5

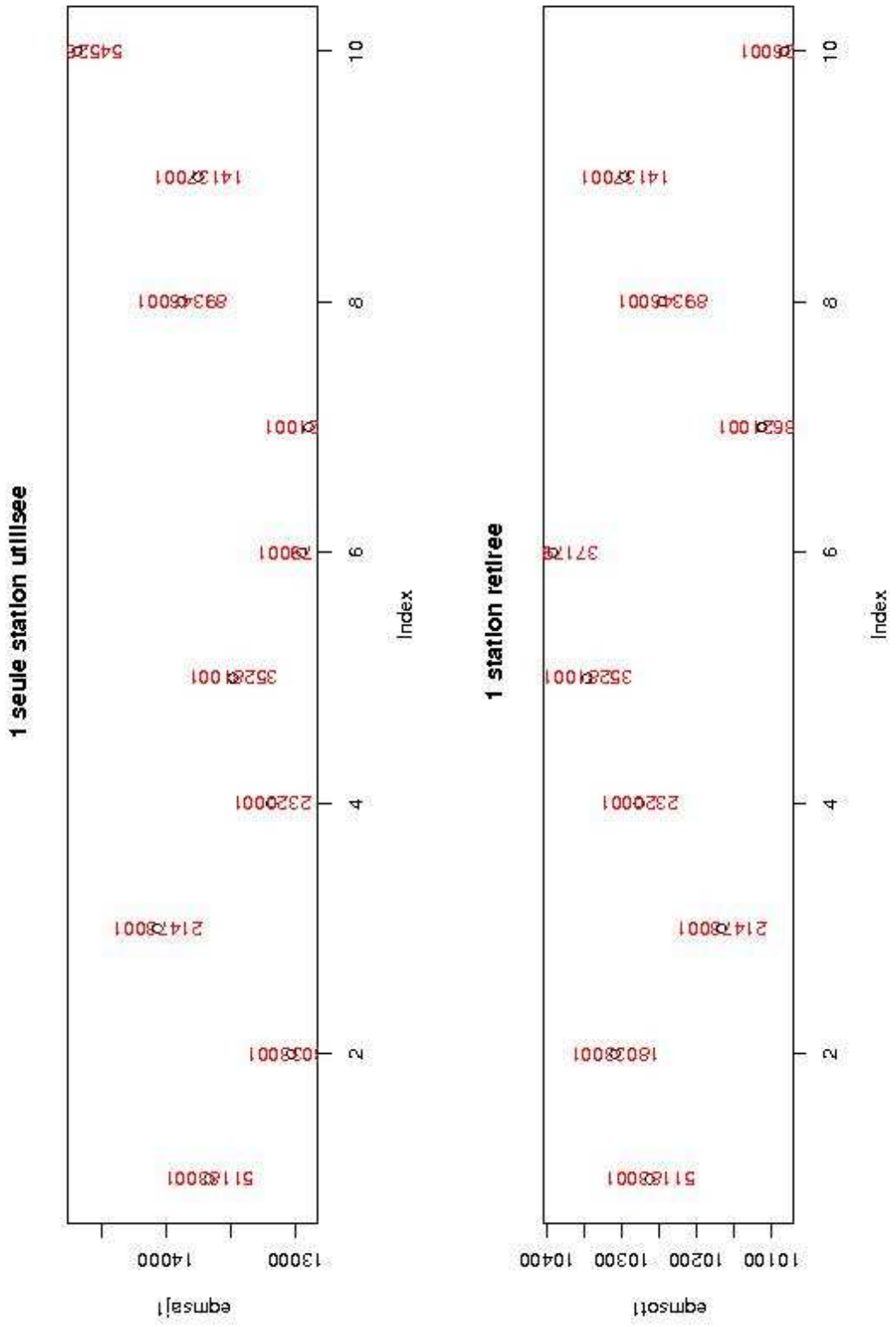
1 seule station utilisee
eqmsalt
13000
14000
Index
2
4
6
8
10
1 station retiree
eqmsott
10100
10200
10300
10400
Index
2
4
6
8
10

FIGURE 6

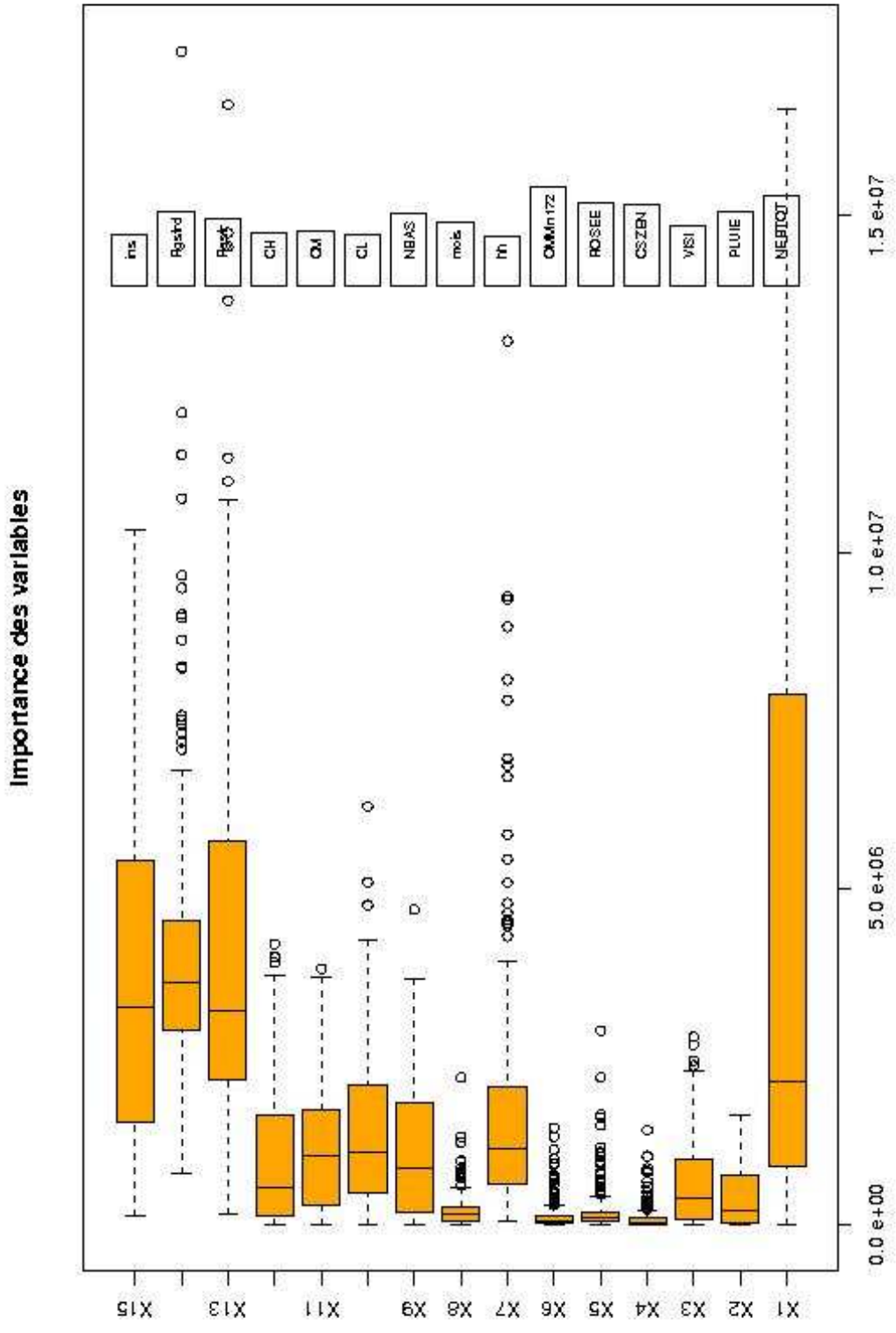
Importance des variables
ins
Rgstrd
Rgstd
CH
CM
CL
NBAS
mois
hh
CMMn172
ROSEE
CSZBN
VISI
PLUIE
NEBTOT
X15
X13
X11
X9
X8
X7
X6
X5
X4
X3
X2
X1
0.0 e+00
5.0 e+06
1.0 e+07
1.5 e+07

FIGURE 7

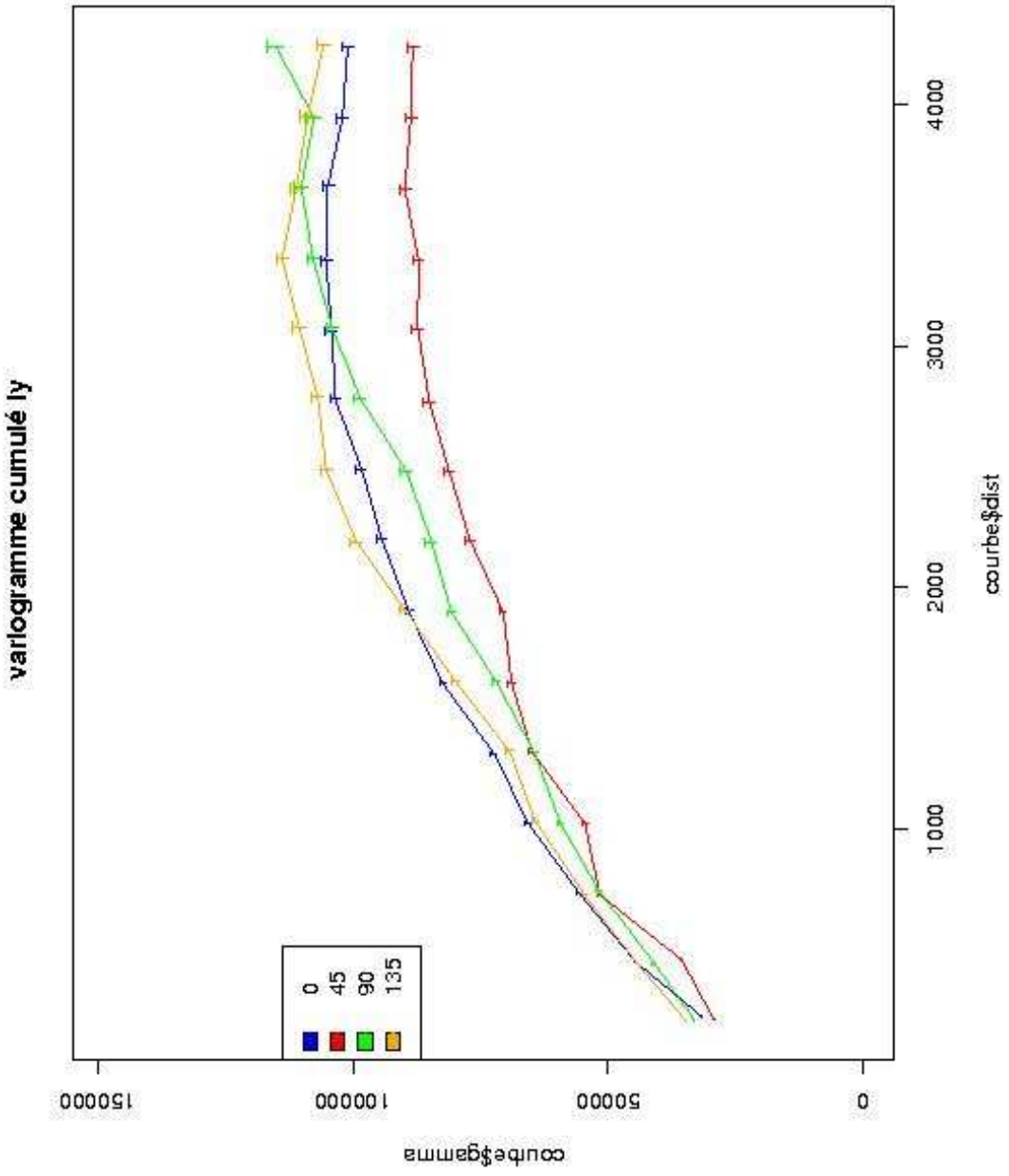
variogramme cumulé ly
courbe$gamma
courbe$dist
0
50000
100000
150000
1000
2000
3000
4000
0
45
90
135

FIGURE 8

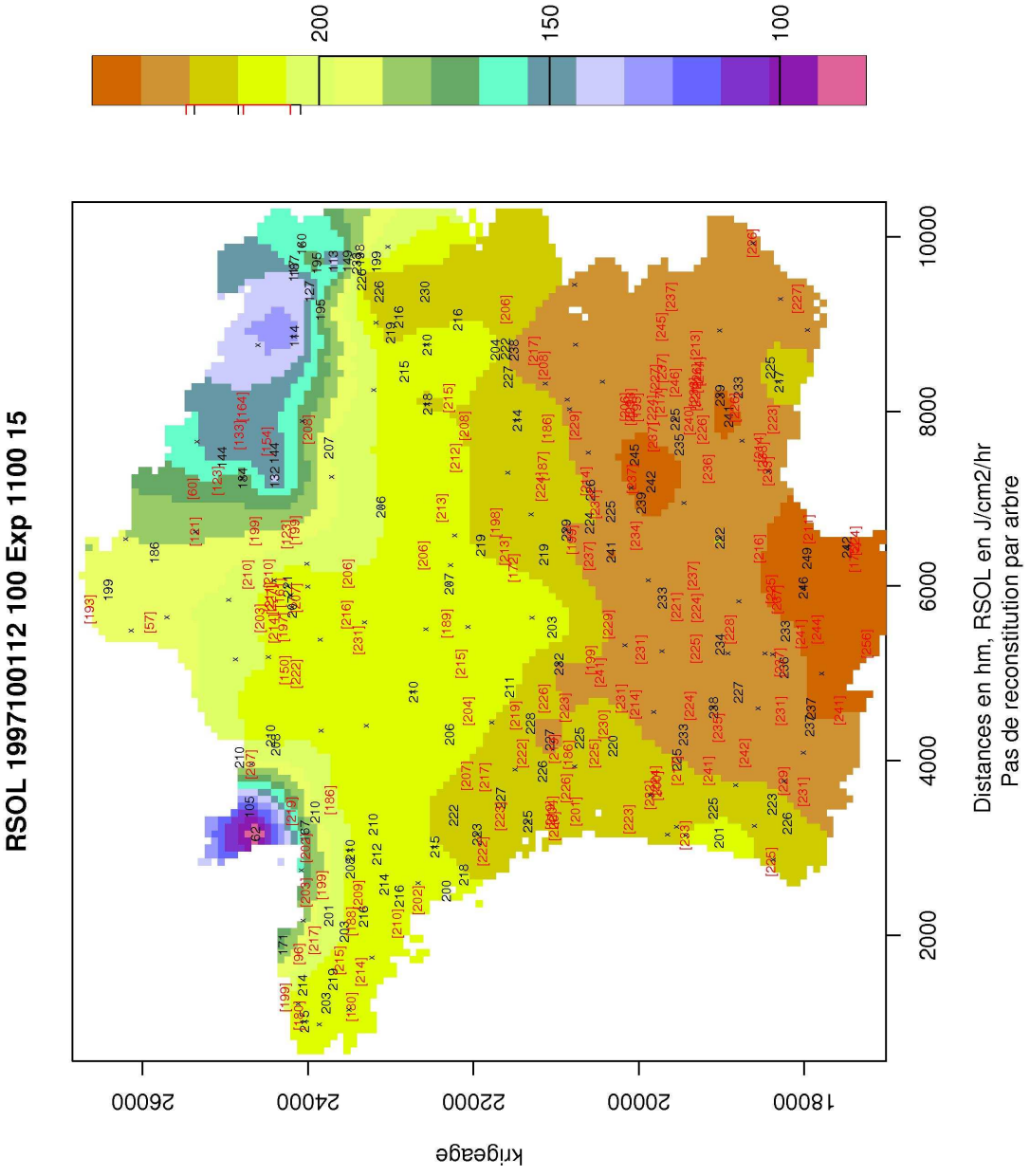
RSOL 1997100112 100 Exp 1100 15
200
150
100
2000
4000
6000
8000
10000
Distances en hm, RSOL en J/cm2/hr
Pas de reconstitution par arbre
krigeage
18000
20000
22000
24000
26000

FIGURE 9

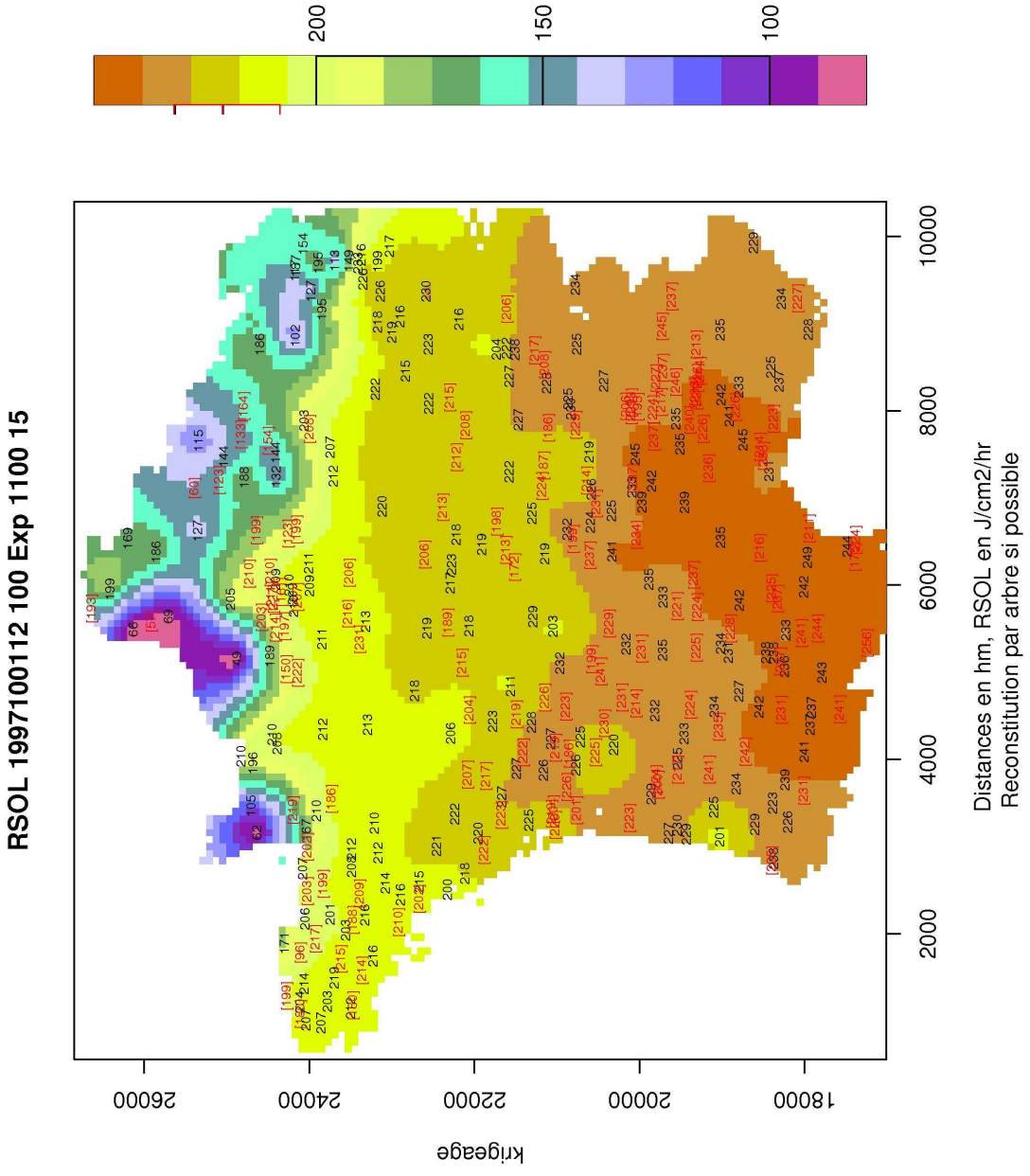
RSOL 1997100112 100 Exp 1100 15
200
150
100
2000
4000
6000
8000
10000
18000
20000
22000
24000
26000
krigeage
Distances en hm, RSOL en J/cm2/hr
Reconstitution par arbre si possible

FIGURE 10

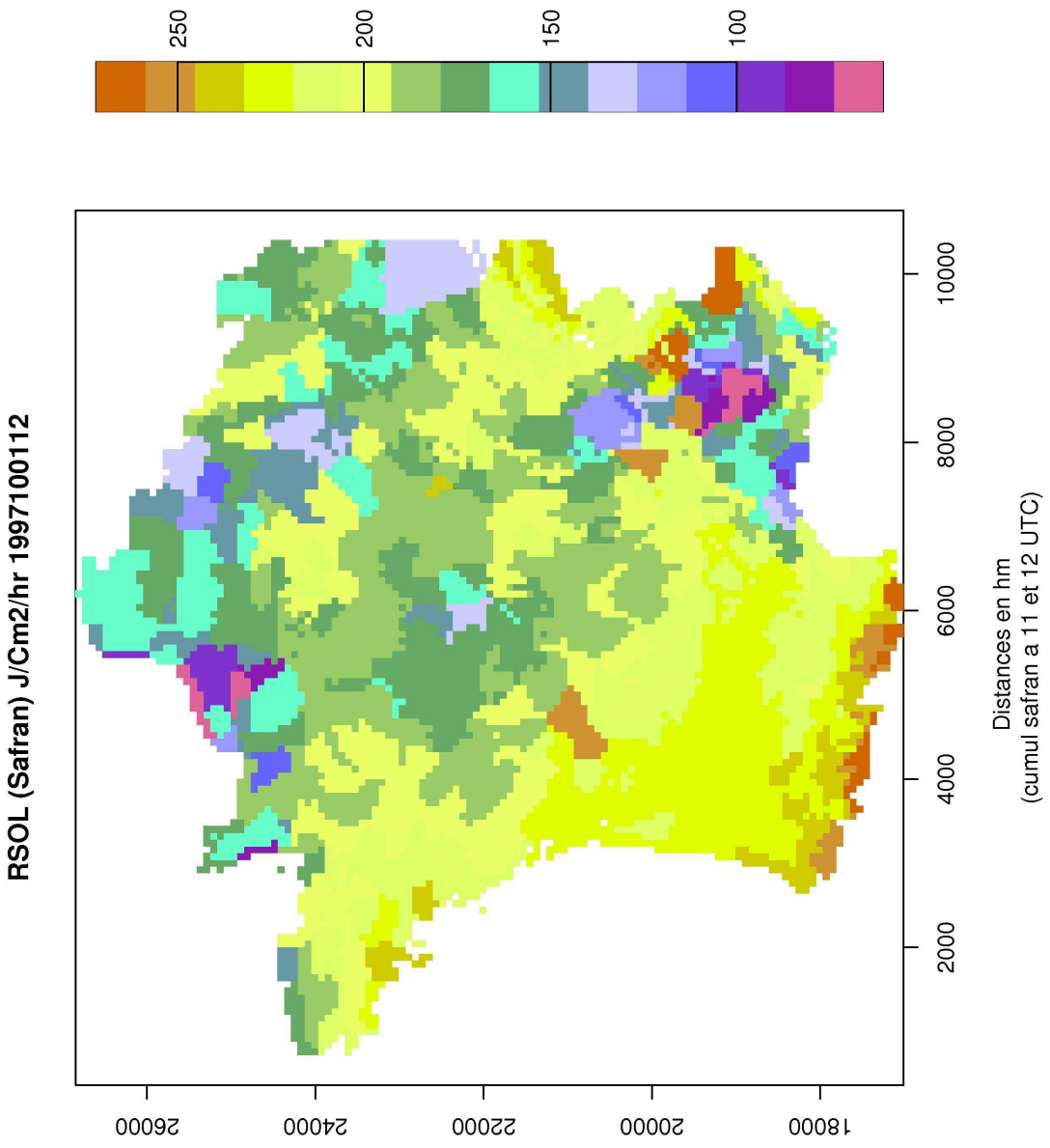
RSOL (Safran) J/Cm2/hr 1997100112
250
200
150
100
18000
20000
22000
24000
26000
2000
4000
6000
8000
10000
Distances en hm
(cumul safran a 11 et 12 UTC)

FIGURE 11

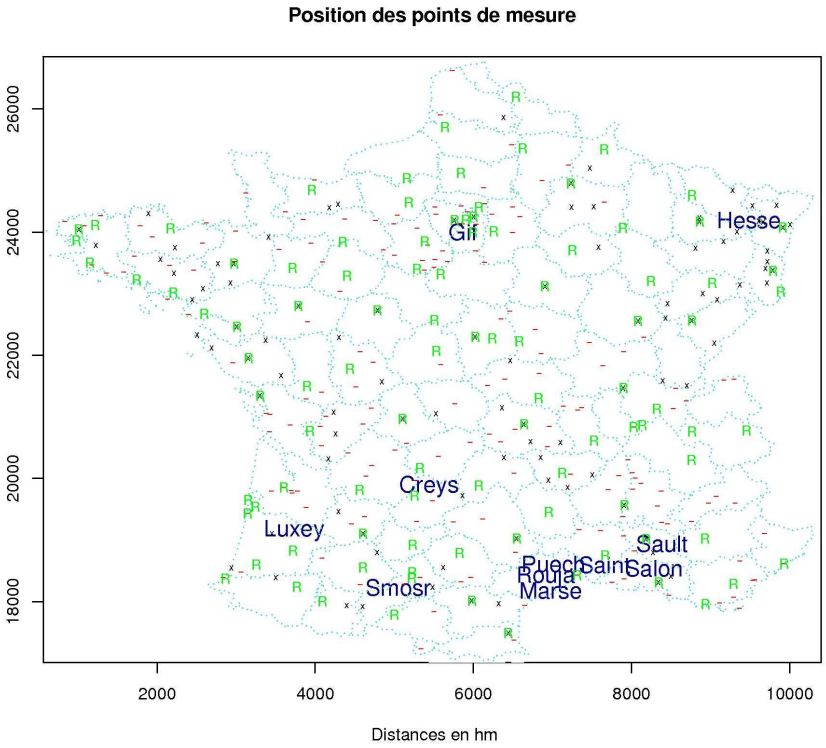
Position des points de mesure
Gif
Hesse
Creys
Luxey
Smosr
Sault
Salon
Marse
26000
24000
22000
20000
18000
2000
4000
6000
8000
10000
Distances en hm

FIGURE 12

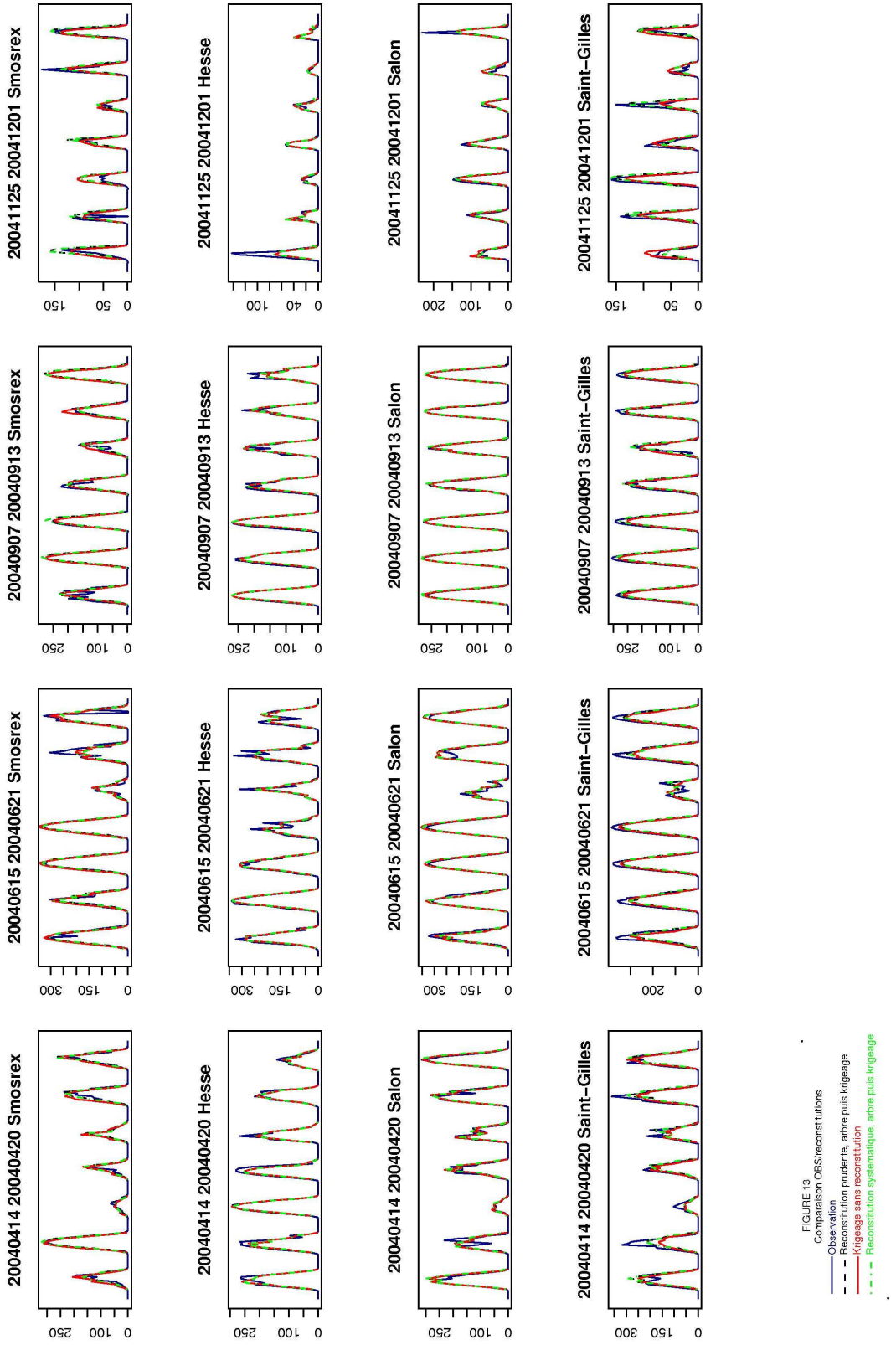
20040414 20040420 Smosrex
20040414 20040420 Hesse
20040414 20040420 Salon
20040414 20040420 Saint-Gilles
20040615 20040621 Smosrex
20040615 20040621 Hesse
20040615 20040621 Salon
20040615 20040621 Saint-Gilles
20040907 20040913 Smosrex
20040907 20040913 Hesse
20040907 20040913 Salon
20040907 20040913 Saint-Gilles
20041125 20041201 Smosrex
20041125 20041201 Hesse
20041125 20041201 Salon
20041125 20041201 Saint-Gilles
FIGURE 13
Comparaison OBS/reconstitutions
Observation
Reconstitution prudente, arbre puis krigeage
Krigeage sans reconstitution
Reconstitution systematique, arbre puis krigeage

FIGURE 13

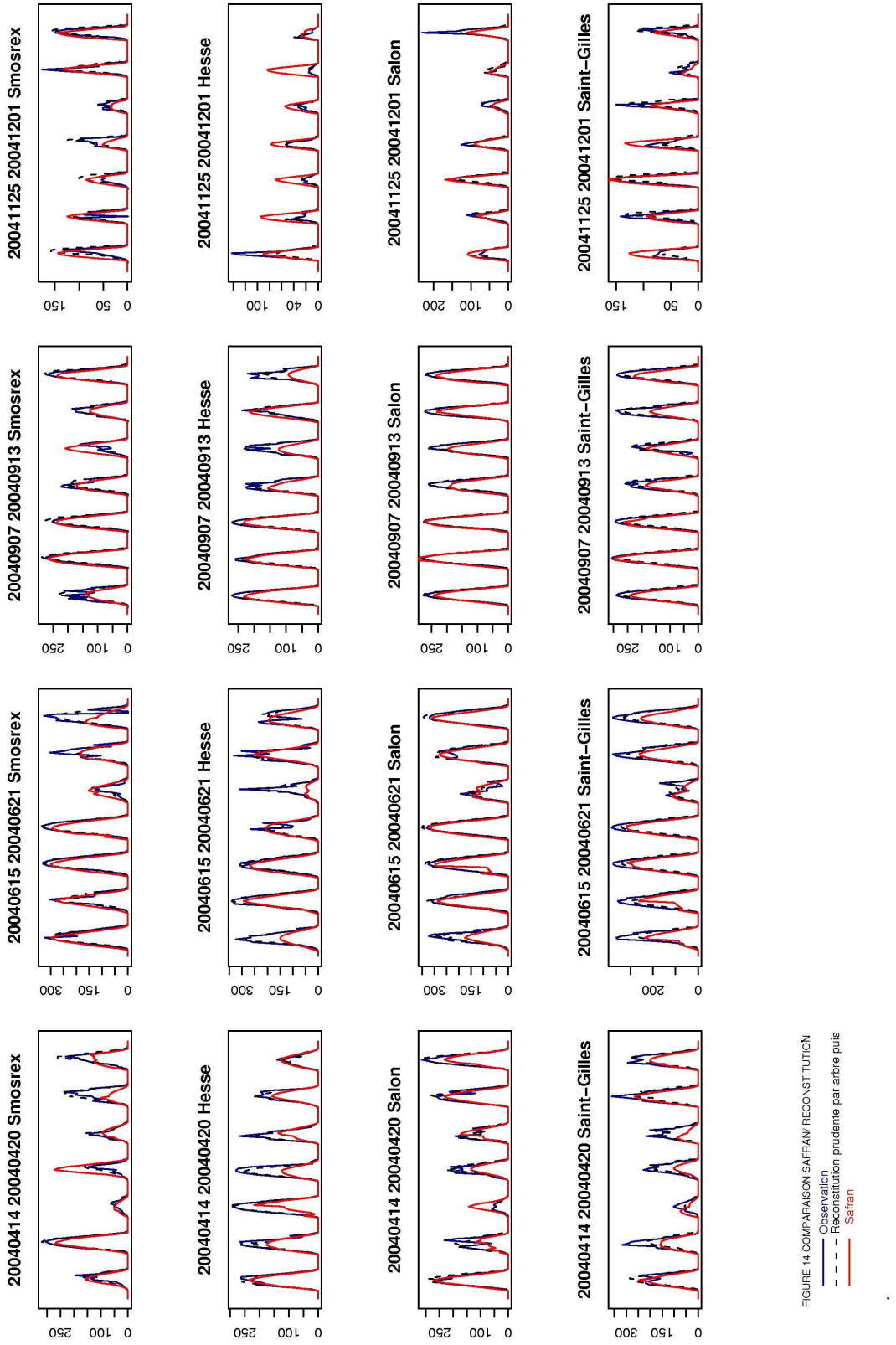
20040414 20040420 Smosrex
20040615 20040621 Smosrex
20040907 20040913 Smosrex
20041125 20041201 Smosrex
20040414 20040420 Hesse
20040615 20040621 Hesse
20040907 20040913 Hesse
20041125 20041201 Hesse
20040414 20040420 Salon
20040615 20040621 Salon
20040907 20040913 Salon
20041125 20041201 Salon
20040414 20040420 Saint-Gilles
20040615 20040621 Saint-Gilles
20040907 20040913 Saint-Gilles
20041125 20041201 Saint-Gilles
FIGURE 14 COMPARAISON SAFRAN/ RECONSTITUTION
Observation
Reconstitution prudente par arbre puis
Safran

FIGURE 14